North East University Bangladesh

Department of Computer Science and Engineering

Improved Image Pre-processing for Better OCR Performance

By

MD Zahid Mahmud Emon
Reg. No: 170103020030
BSc(Engg) in CSE
4th year 3rd semester

Gulam Kibria Chowdhury
Reg. No: 170103020033
BSc(Engg) in CSE
4th year 3rd semester

Supervised By

Tasnim Zahan
Assistant Professor
Department of Computer Science and Engineering

17th February, 2021

Improved Image Pre-processing for Better OCR Performance

A Thesis submitted to the Department of Computer Science and Engineering,
North East University Bangladesh, in partial fulfillment of the requirements
for the degree of Bachelor of Science in Computer Science and Engineering

By

MD Zahid Mahmud Emon
Reg. No: 170103020030
BSc(Engg) in CSE
4th year 3rd semester

Gulam Kibria Chowdhury
Reg. No: 170103020033
BSc(Engg) in CSE
4th year 3rd semester

Supervised By

Tasnim Zahan
Assistant Professor
Department of Computer Science and Engineering

17th February, 2021

Recommendation Letter from Thesis Supervisor

These Students, Gulam Kibria Chowdhury & Md Zahid Mahmud Emon, whose thesis entitled *"Improved Image Pre-processing for Better OCR Performance",* is under my supervision and agrees to submit for examination.

Signature of the Supervisor :

Tasnim Zahan
Assistant Professor
Department of Computer Science and Engineering
North East University Bangladesh

Qualification Form of BSc(Engg.) Degree

Students Name : Gulam Kibria Chowdhury & Md Zahid Mahmud Emon
Thesis Title : Improved Image Pre-processing for Better OCR Performance

This is to certify that the thesis is submitted by Gulam Kibria Chowdhury & Md Zahid Mahmud Emon in February, 2021. It is qualified and approved by the following persons and committee.

Head of the Dept.

Tasnim Zahan
Assistant Professor
Department of CSE
North East University Bangladesh

Supervisor

Tasnim Zahan
Assistant Professor
Department of CSE
North East University Bangladesh

Abstract

Documents of historical significance have very limited use in this digital era. Handwritten documents are decreasing day by day. Images of those documents are the only thing that matters now. But degraded documents have very noisy images. Which makes the preservation process a bit difficult. Also Different size images makes it difficult to build a system which can clean the images and also preserve the information of those images. Therefore, OCR doesn't perform well for those images. We are using a technique for pre-processing variable size noisy document images, which will improve the OCR performance significantly. Our approach focuses on cleaning the noisy images while not losing any original information from the document images. We tested several older methods and came up with a newer variation for cleaning images and preserving its information. Handling different size of images, we used several techniques like Zero Padding, Image cropping and for cleaning, we used CNN-Auto-encoder and combining these approach we cleaned and preserved all information within images.

Keywords: Image pre-processing, OCR, Degraded Document Images, CNN, Auto-encoder, Noise, Image Cropping, Zero Padding, Image Reconstruction.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Image processing for document images is a widely researched topic. Image processing is a very important step for improved OCR (Optical character Recognition) performance. OCR is used for various modern world problems, so in order to solve those problems much developed systems have to be in place. But to develop such systems image processing plays a very important role. Many researches had been conducted to develop those systems. And the research in this topic is increasing day by day. Image processing is used for extracting some useful information in an image or just for enhancing the image. It takes an input image and gives output of an image or some useful data about the image. That information about the image can be used in OCR for better results. Image pre-processing step is actually used as image processing in OCR. Image pre-processing is a well-known technique for enhancing the image and removing any kind of distortion from an image. So image pre-processing for document images and also captured images are very useful for better OCR. We live in an era where hand-written documents are replaced by digital documents. But to preserve those documents and their information we need a quick and useful solution. But when we try to collect and preserve those documents in digital form we face some terrible issues like noise and degradation in documents. A combination of image pre-processing and OCR can remove these issues for the rescue.

Though Image pre-processing is a well-documented field, our key idea is to improve even more in this field by introducing new ideas for image pre-processing and its use case.

The rest of this chapter is about Image processing, Image pre-processing, OCR.

1.1 Image Processing

Image processing is the technique used for manipulation of images for various reasons. Its use is increasing day by day. Its use case is spread out in many scientific fields like medicine, entertainment, geology, architecture etc. From our everyday used tools like mobile to camera everywhere we see the use of image processing. Multimedia systems, one of the most valued systems, is based on image processing. Image processing can be separated into many classes but most significant ones being image enhancement, image restoration, image analysis, image compression.

- **Image Enhancement:** Image enhancement is a technique by which an image is manipulated so that a user or a viewer can extract more information out of that image. It is very well-known as an OCR research field.
- **Image Restoration:** Image restoration technique is used in restoration of corrupted images where it traces back the degradation data so that it can revert the whole process to get back the original image.
- **Image Analysis:** This technique is used for extracting information automatically from processed images. Image segmentation, edge extraction texture and motion analysis are some of the image analysis steps.
- **Image Compression:** We need a huge amount of data to represent an image. A normal quality gray scale image of n*n needs n*n*8 bits for representation. So storing and transmitting this huge data we need to compress it to our comfort. The redundancy of information is exploited for reducing the number of bits in the image.
- **Image Cropping:** This technique is used for removing any unnecessary areas of an image. The process usually consists of the removal of some of the peripheral areas of an Image to remove extra un-useful space for improve image framing and aspect ratio. This technique is used on artwork, captured digital images.
- **Zero Padding:** It is a used in convolutional neural network as it refers to the amount of pixels added to an image when it is being processed by the kernel of an CNN.

1.2 Image Pre-processing:

Image pre-processing is a very well documented topic in modern research. Image pre-processing is used for the initial step for various research like computer vision and OCR.
Image pre-processing can be a simple task like image resizing or converting the image color to grayscale from RGB. Other pre-processing like geometrical shape transformation of an image is also used in research.

Image pre-processing can be described as these steps--

1) Images are taken as input
2) Padding or cropping is done if needed
3) Converted into RGB to Grayscale

4) Removal of any degradation in the image
5) Returns more useful output image

1.2.1 Use of Image Pre-processing

Document images containing text or graphics cannot be in colored format, it needs to be in grayscale format or binary image format since the colored images needed high computational resources. They also may have non-uniform backgrounds, which makes the text extraction process very difficult for document images. To solve all these problems, we need pre-processing. The steps are being first we need to enhance the image by removing any kind of noise and degradation and correcting the contrast in the image. Then if needed we need to correct the color format or just convert the image into binary image by using the thresholding technique to remove the background containing any scenes and watermarks. These steps help an image to be improved while maintaining its original and important information, which later can be used in OCR or computer vision problems.

The acquired data from different sources vary in quality and material, so when we need to feed these data to an advanced network or system they need to be standardized. Pre-processing reduces the complexity and increases the accuracy of the applied algorithm. We can't always write or build unique models for different images, so we just acquire an image and convert it such a way that the system allows the data to solve a problem.

Modern research in OCR or computer vision requires more pre-processing steps than any other research field. So in completion of those research fields and their result well planned and well-designed pre-processing is much needed.

1.2.2 How Image pre-processing works?

Image pre-processing steps differ in different types of problems. But the fundamental functionality is the same for all. Some of the problems just need re-scaling, some just need the color correction and some just may need noise removal.

Some of the Image pre-processing technique is given below:

- **Color Conversion:** Some of the modern research problems just only need color correction or color conversion in the pre-processing section. Contrast of the image and colored images sometimes needs more computational power than any other images.

By color correction of the input image we can reduce the required computational resources. Same goes for color conversion problems. Because colored image processing needs a highly configured GPU and many other resources but we can ease the load by converting the colored image into a grayscale image.

- **Re-scaling:** In some of the research problems we need to adjust our image according to the prepared model. Adjusting an image can mean enlarging the image or scaling down the image, most of the time the images need to be scaled down for better training. But enlargement of the image is also necessary when we need to extract more valuable information from an image. Re-scaling means re-sizing the original raster image into scaled down or scaled up new image.

- **Blurring:** Image blurring is usually achieved by connecting the image with a low-pass filter kernel. Filters are usually used to blur the image or reduce noise but there are some exceptions and differences between them.

  Some of the blurring method are -

  1) **Averaging:** After connecting the image with a normalize box of filter, this simply takes the average of all the pixels under the kernel area and replaces the central element.

  2) **Gaussian blurring:** This works as same as averaging but uses Gaussian kernel. Here, the dimensions of the kernel and standard deviation in both direction can be determined independently. Gaussian blurring is very useful for removing Gaussian noise. And it does not preserve the edges of the input.

  3) **Median blurring:** The central element in the kernel area is replaced with the median of all the pixels under the kernel. This method outperforms other blurring method in removing the famous salt-pepper noise from the images. Median filter is a non-linear filter. Median blurring replaces the pixel values with median values available in the neighborhood values. Median values preserves the edges of the input

- **Thresholding:** There is not a single image thresholding method that fits all types of documents. All filters perform differently on varying images.

Here are some thresholding method -

1) **Simple threshold:** By the name it reminds us about simple thresholding. A single threshold value is set to determine whether a pixel will become black or white. It also goes with the name Binary thresholding.

2) **Adaptive threshold:** Rather than setting a one global threshold value the algorithm calculate the threshold for small regions of the image. So, system end up aveing various threshold values for different regions of the image. There are two adaptive methods for calculating the threshold value.
    I. **Adaptive Thresh Mean**
    II. **Adaptive Gaussian Mean**

3) **Otsu's threshold:** This method works well with **Bimodal images**, which is an image whose histogram has two peaks. We pick the threshold value between these peaks. This method does not do well other than **Bimodal images**.

1.3 OCR (Optical Character Recognition)

OCR (Optical Character Recognition) is a widespread technology to recognize text inside images, such as scanned documents and photos. OCR technology is used to convert any kind of images containing written text (typed, handwritten or printed) into machine-readable text data.

OCR Technology became famous in early 1990s while attempting to digitize historic newspapers. Since then the technology is gone through several improvements.

1.3.1 Image Pre-processing in OCR

OCR software system often pre-processes images to improve the chances of successful recognition. The aim of image pre-processing is to improve the actual image data. By doing so, unwanted distortions are eliminated and specific feature of the image are enhanced.

1.3.2 Use Case of OCR

The most well-known and used OCR field is converting printed documents into machine readable text documents. Once a scanned document paper went through OCR processing, the text of the document can be edited with word processor like Google docs. Some years ago the only option to digitize printed paper documents was to manually re-typing the text. Which was massively time consuming and most of the cases inaccurate.

OCR is secretively often used in many well-known systems and services in our life. The most important ones being indexing documents for search engines, automatic number plate recognition as well assisting blind and visually impaired persons.
OCR technology has also proven significantly useful in digitizing historic newspapers and texts that have now been converted into fully searchable formats and has made accessing those earlier texts easier and faster.

The possibilities for using OCR software is widely useful. Because OCR can be combined with a broad range of technologies. Identification process in OCR is one of the examples of this category. All identity cards like passports and IDs have machine readable zones that can be scanned, OCR can speed up this process of identifying and registering people. Border checkpoints security forces can use this method very easily.

OCR brings more than enough opportunity in our life with various use cases of its own offering.

1.4 CNN-Autoencoder

An Autoencoder is a neural network that learns to copy its input to its output. The internal mechanism of coping its input to output is completely relied on two main parts: an encoder that maps the input structure to a code and a decoder that reconstruct the input through that map and produce the output.

Denoising autoencoders are used to achieve good representation of the input by changing the reconstruction conditions. This type of autoencoder can be applied for any kind of noises. Some examples being additive isotropic Gaussian noise, Masking noise or salt paper noises.
The noise adding of the input is performed only during training. Once the model learns the optimal parameters, in order to extract the representations from the original data no noising work need to be done.

A convolutional auto-encoder is a neural network that is trained to reproduce its input image in the output layer. This process is combination of the encoding part and decoding part, where encoder just compresses the image by producing a low dimensional representation of the input image and decoder takes the compressed image and reconstruct the image with only the important data.

1.5 Problem Statement

Without pre-processing OCR engines returns not so promising results. Noise, shadows and contrast of the image causes OCR engines to Give bad results. Pre-processed data improves the performance of any OCR in a significant rate.

We are working on degraded document images for preserving its original information. We have tried the preserving process by sending the documents without any pre-processing to existing OCR engine like 1. Shulikhan OCR (SUST) and 2. Tesseract OCR (Open Source).
But both of them returns not so promising results.

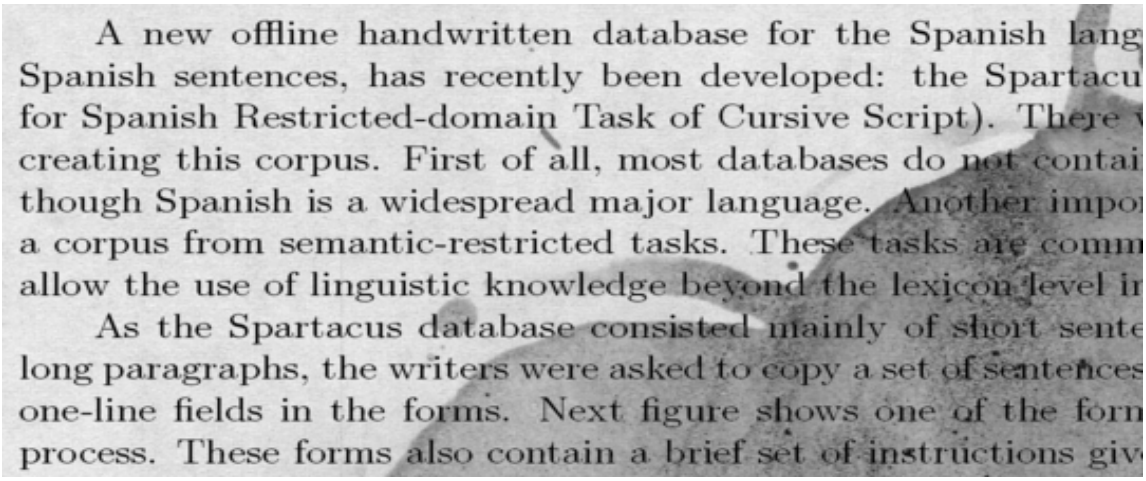

A new offline handwritten database for the Spanish lang
Spanish sentences, has recently been developed: the Spartacu
for Spanish Restricted-domain Task of Cursive Script). There
creating this corpus. First of all, most databases do not contai
though Spanish is a widespread major language. Another impo
a corpus from semantic-restricted tasks. These tasks are comm
allow the use of linguistic knowledge beyond the lexicon level i
As the Spartacus database consisted mainly of short sente
long paragraphs, the writers were asked to copy a set of sentences
one-line fields in the forms. Next figure shows one of the form
process. These forms also contain a brief set of instructions giv

Figure 1: Input image 1 [12]

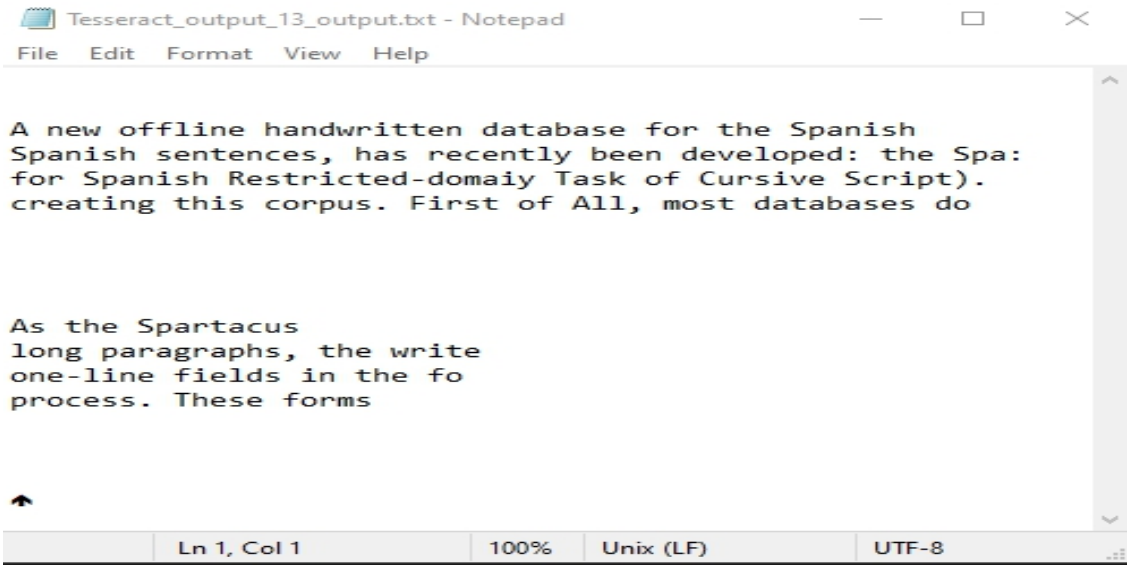
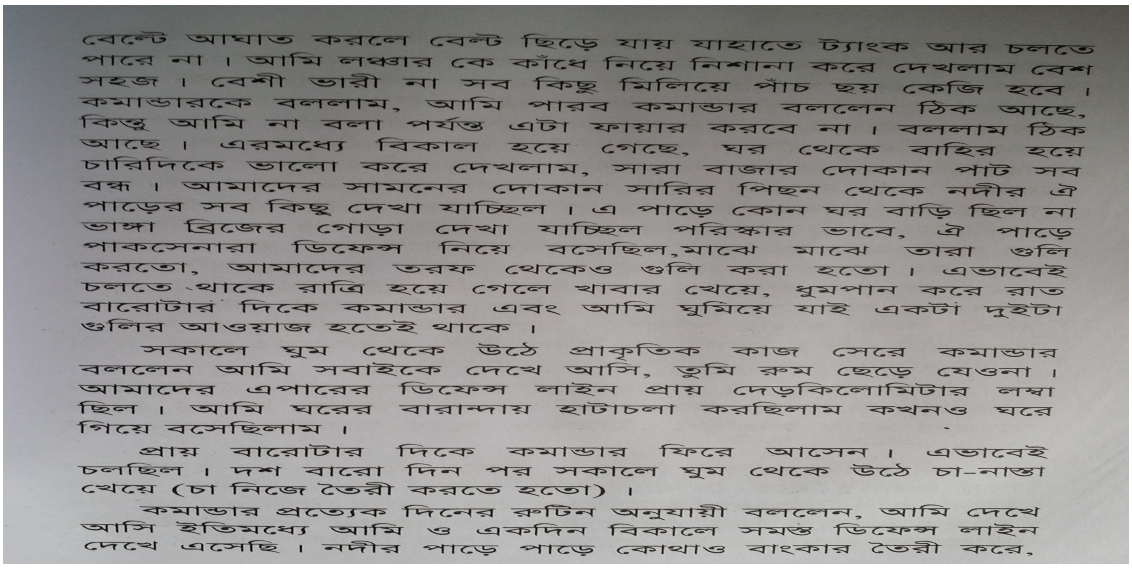


Figure 2: Tesseract OCR output text file image

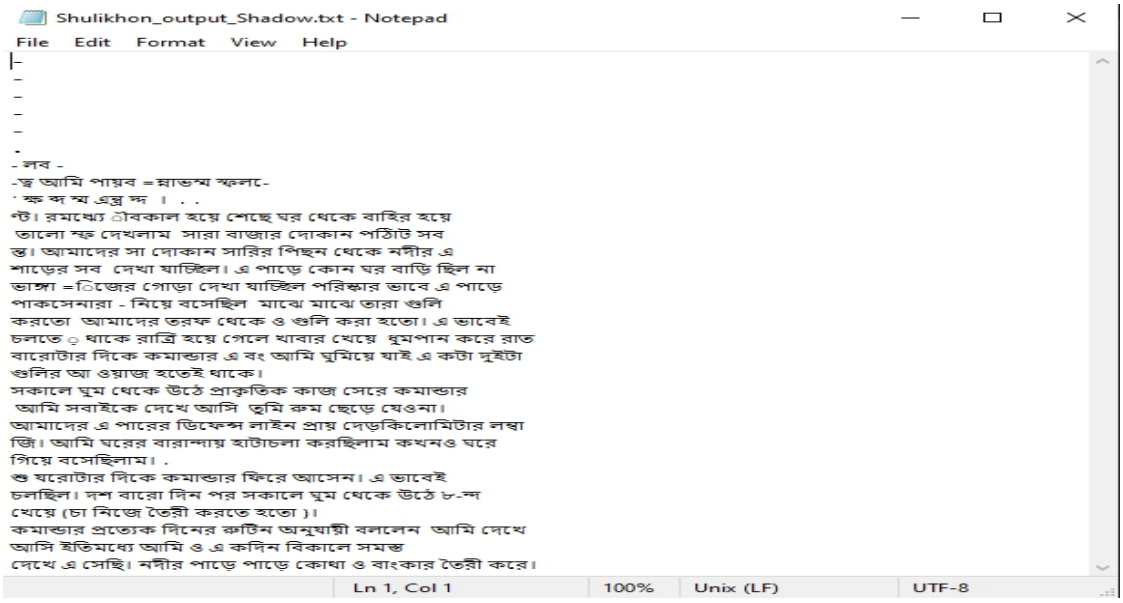

বেল্টে আঘাত করলে বেল্ট ছিড়ে যায় যাহাতে ট্যাংক আর চলতে পারে না । আমি লঞ্চার কে কাঁধে নিয়ে নিশানা করে দেখলাম বেশ সহজ । বেশী ভারী না সব কিছু মিলিয়ে পাঁচ ছয় কেজি হবে । কমান্ডারকে বললাম, আমি পারব কমান্ডার বললেন ঠিক আছে, কিন্তু আমি না বলা পর্যন্ত এটা ফায়ার করবে না । বললাম ঠিক আছে । এরমধ্যে বিকাল হয়ে গেছে, ঘর থেকে বাহির হয়ে চারিদিকে ভালো করে দেখলাম, সারা বাজার দোকান পাট সব বন্ধ । আমাদের সামনের দোকান সারির পিছন থেকে নদীর ঐ পাড়ের সব কিছু দেখা যাচ্ছিল । এ পাড়ে কোন ঘর বাড়ি ছিল না ভাঙ্গা ব্রিজের গোড়া দেখা যাচ্ছিল পরিষ্কার ভাবে, ঐ পাড়ে পাকসেনারা ডিফেন্স নিয়ে বসেছিল,মাঝে মাঝে তারা গুলি করতো, আমাদের তরফ থেকেও গুলি করা হতো । এভাবেই চলতে থাকে রাত্রি হয়ে গেলে খাবার খেয়ে, ধুমপান করে রাত বারোটার দিকে কমান্ডার এবং আমি ঘুমিয়ে যাই একটা দুইটা গুলির আওয়াজ হতেই থাকে ।

সকালে ঘুম থেকে উঠে প্রাকৃতিক কাজ সেরে কমান্ডার বললেন আমি সবাইকে দেখে আসি, তুমি রুম ছেড়ে যেওনা । আমাদের এপারের ডিফেন্স লাইন প্রায় দেড়কিলোমিটার লম্বা ছিল । আমি ঘরের বারান্দায় হাটাচলা করছিলাম কখনও ঘরে গিয়ে বসেছিলাম ।

প্রায় বারোটার দিকে কমান্ডার ফিরে আসেন । এভাবেই চলছিল । দশ বারো দিন পর সকালে ঘুম থেকে উঠে চা-নাস্তা খেয়ে (চা নিজে তৈরী করতে হতো) ।

কমান্ডার প্রত্যেক দিনের রুটিন অনুযায়ী বললেন, আমি দেখে আসি ইতিমধ্যে আমি ও একদিন বিকালে সমস্ত ডিফেন্স লাইন দেখে এসেছি । নদীর পাড়ে পাড়ে কোথাও বাংকার তৈরী করে,

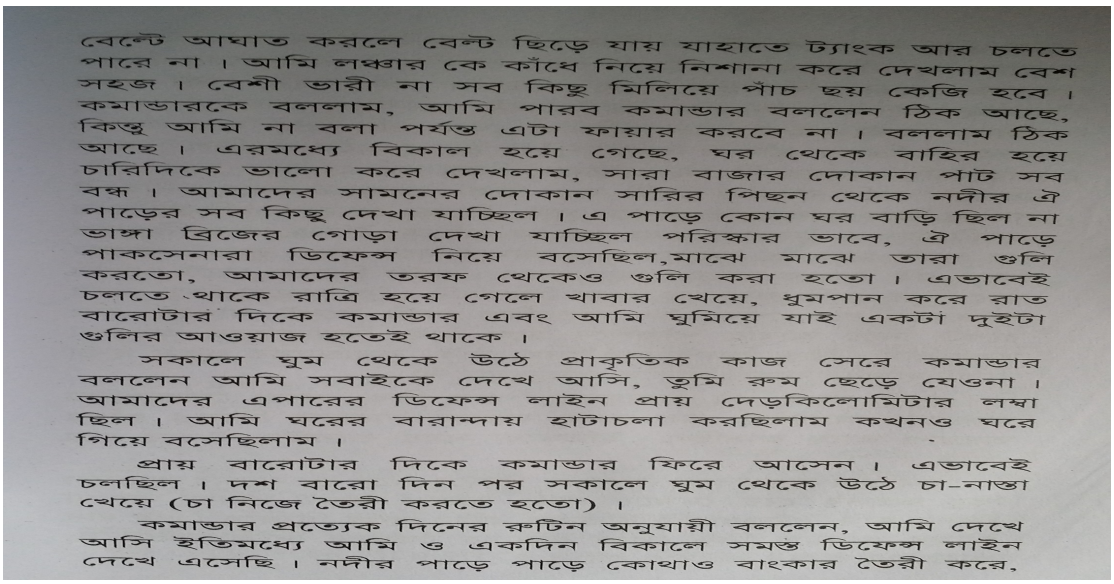
Figure 3: Input image 2


Shulikhon_output_Shadow.txt - Notepad

File Edit Format View Help

Ln 1, Col 1 100% Unix (LF) UTF-8

Figure 4: Shulikhan OCR output test file image


বেল্টে আঘাত করলে বেল্ট ছিড়ে যায় যাহাতে ট্যাংক আর চলতে পারে না । আমি লঞ্চার কে কাঁধে নিয়ে নিশানা করে দেখলাম বেশ সহজ । বেশী ভারী না সব কিছু মিলিয়ে পাঁচ ছয় কেজি হবে । কমান্ডারকে বললাম, আমি পারব কমান্ডার বললেন ঠিক আছে, কিন্তু আমি না বলা পর্যন্ত এটা ফায়ার করবে না । বললাম ঠিক আছে । এরমধ্যে বিকাল হয়ে গেছে, ঘর থেকে বাহির হয়ে চারিদিকে ভালো করে দেখলাম, সারা বাজার দোকান পাট সব বন্ধ । আমাদের সামনের দোকান সারির পিছন থেকে নদীর ঐ পাড়ের সব কিছু দেখা যাচ্ছিল । এ পাড়ে কোন ঘর বাড়ি ছিল না ভাঙ্গা ব্রিজের গোড়া দেখা যাচ্ছিল পরিষ্কার ভাবে, ঐ পাড়ে পাকসেনারা ডিফেন্স নিয়ে বসেছিল,মাঝে মাঝে তারা গুলি করতো, আমাদের তরফ থেকেও গুলি করা হতো । এভাবেই চলতে থাকে রাত্রি হয়ে গেলে খাবার খেয়ে, ঘুমপান করে রাত বারোটার দিকে কমান্ডার এবং আমি ঘুমিয়ে যাই একটা দুইটা গুলির আওয়াজ হতেই থাকে ।

সকালে ঘুম থেকে উঠে প্রাকৃতিক কাজ সেরে কমান্ডার বললেন আমি সবাইকে দেখে আসি, তুমি রুম ছেড়ে যেওনা । আমাদের এপারের ডিফেন্স লাইন প্রায় দেড়কিলোমিটার লম্বা ছিল । আমি ঘরের বারান্দায় হাটাচলা করছিলাম কখনও ঘরে গিয়ে বসেছিলাম ।

প্রায় বারোটার দিকে কমান্ডার ফিরে আসেন । এভাবেই চলছিল । দশ বারো দিন পর সকালে ঘুম থেকে উঠে চা-নাস্তা খেয়ে (চা নিজে তৈরী করতে হতো) ।

কমান্ডার প্রত্যেক দিনের রুটিন অনুযায়ী বললেন, আমি দেখে আসি ইতিমধ্যে আমি ও একদিন বিকালে সমস্ত ডিফেন্স লাইন দেখে এসেছি । নদীর পাড়ে পাড়ে কোথাও বাংকার তৈরী করে,

Figure 5: Input image 2

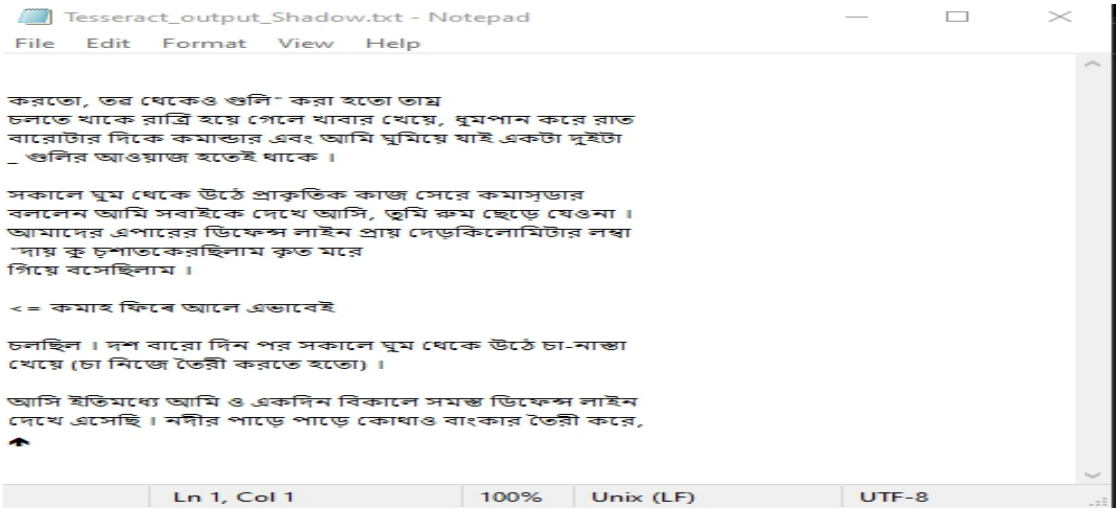


Figure 6: Tesseract OCR output text file image

Now, we are trying to find a solution for preserving the document images with necessary pre-processing step to better OCR performance.

1.6 Outline of the report

The first chapter of this report information the basics of image processing, image pre-processing, OCR and CNN-Autoencoder which is required for our approach.

In second chapter of this report some articles related to image pre-processing for OCR performance improvement have been summarized and these articles elaborates some key approach for image pre-processing.

In third chapter of this report, we have described the preprocessing steps for images and described our collected and created datasets, cleaning process of the images. We also described our Methodology of this work.

In fourth chapter of the report includes the result of the work is described. We described the accuracy as well as some analysis and comparison of our work.

In the last chapter we have summarized our work and conclusion of the work done.

Chapter 2

BACKGROUND STUDY

There have been many research done for better OCR performance using image processing and pre-processing. We have studied and chosen some of the top research where image pre-processing is done for OCR. This chapter is about all the related work that have been done in betterment of OCR using image pre-processing.

2.1 Selecting automatically pre-processing methods to improve OCR performance.

In research [1] Quang Anh BUI, David MOLLARD and Salvatore TABBONE proposed an approach that automatically selects suitable document pre-processing algorithm to increase OCR performance. They first provided experimental evaluation of different pre-processing methods on different OCR engines for document image of different distortions. In the context that distortions on the document and information about OCR system's are unknown they proposed an automatic pre-processing selection method based on convolutional neural network with 15 layers and where the last layer contains neurons representing their different pre-processing algorithms. Effectiveness of their approach to improve OCR performance shows in the experimental results for mobile captured document images.

Image set 1contains 1000 blurry images
Set 2 contains 1000 noisy images and
Set 3 contains 1000 balanced images. Random combination of noisy, blurry and good quality images
Here are the results of the performance of OCR with the proposed selection method -

OCR systems	Tesseract (v2.0)	NNOCR
Set 1	0.54	0.88
Set 2	0.40	0.95
Set 3	0.04	0.90

Table 1: Performance of OCR systems with the proposed selection method.

Experimental results shows the efficiency of their CNN-based method. Because the performance of each OCR are improved after the automatic pre-processing methods selection.

Performance of OCR systems with or without the proposed selection method-

OCR systems	Tesseract (v2.0)	NNOCR
Without selection method	0.38	0.95
With selection method	0.63	0.96

Table 2: Performance of OCR systems with or without the proposed selection method.

2.2 Efficient binarization technique for severely degraded document images

In research [2] Brij Mohan Singh and Mridula proposed a way to overcome the degradation problem of document images by a robust binarization technique that recovers the text from a severely degraded document image and thereby increase the accuracy of OCR (Optical Character Recognition) systems. Proposed work is a fusion of two well-known binarization methods: Gatos et al. And Niblack, using dilation and logical AND operations. This fusion gives far better results than the existing binarization methods.

The reason behind combining Niblack approach with Gatos et al. Is Niblack method can distinguish the text from the background in the area close to the text. And it can completely recovers the text from badly degraded images. The fusion of these two method complements each other on their weak points.

Here are the evaluation results of the proposed method comparing the traditional binarization method-

Method	Time (s)	F-measure (%)	PSNR (dB)	NRM (10-2)
Otsu	0.017	36.80	07.68	27.21
Gatos	9.53	72.63	39.78	13.28
Niblack	02.83	47.92	17.26	14.35
Sauvola	02.92	42.38	29.34	32.29
Bernsen	02.87	54.32	21.09	11.21
Proposed	12.34	90.42	40.42	7. 59

Table 3: Evaluation results

2.3 OCR Accuracy improvement on document images through a novel pre-processing approach.

In research [3] A. El Harraj and N. Raissouni proposed a novel nonparametric and unsupervised method to compensate for undesirable document image distortions aiming to optimally improve OCR accuracy. Proposed method relies on a very efficient stack of document image enhancing techniques to recover deformation of the entire document image. To solve the lighting variation and irregular distribution of image illumination they proposed a local brightness and contrast adjustment method. They used an optimized grayscale conversion method to transform the document image into grayscale image. They used the Unsharp Masking method to sharpen the grayscale image. Lastly, they used an optimal global binarization technique to prepare the final document image for OCR. This four step novel approach for pre-processing significantly improved text detection rate and optical character recognition accuracy.

Here are the results of the novel pre-processing approach with respect to the original image is below-

Document ID	Original image		Processed image	
	Errors	Accuracy	Errors	Accuracy
AR00233	10	99.34%	0	100%
AR00235	60	99.01%	30	99.5%
AR00270	312	77.17%	219	83.97%
AR00319	89	97.64%	15	99.6%

Table 4: Character accuracy using Original and processed document image

2.4 Document Preprocessing System - Automatic Selection of Binarization

In research [4] Ines Ben Messaoud, Hamid Amiri. Haikal El Abed and Volker Margner proposed a system which allows automatic preprocessing of historical documents. The proposed system is applied on a set of books from the Google-Books (23 books,1000 images) and the Bayerische Staatsbibliothek (10 books, 750 images) collection. One of many preprocessing methods as well as sets of input parameters are selected for each book from the used database according to the input image features. Testing is done on every step during the training and validation of the carried results is performed on another subset of images.

Results of the evaluation of the selected methods on a new subset of documents form BSB and GOOGLE-BOOKS collection is below-

		F M	P-FM	PSNR	N RM	MPM	GA	ρ
		(%)	(%)		·10−2	·10−3	·102	·102
BSB	1st method	86.48	86.58	16.43	5.48	3.32	92.09	85.49
	2nd method	86.58	86.54	16.18	5.51	4.79	92.95	85.36
	3rd method	88.65	89.05	16.94	5.30	2.74	93.97	87.49
Google-Books	1st method	86.72	87.56	16.57	3.14	3.88	95.59	86.13
	2nd method	85.61	86.12	16.1	2.5	6.97	96.07	85.1
	3rd method	89.55	90.91	17.81	3.42	6.32	95.63	88.83

Table 5: Evaluation of the selected methods on a new subset of documents from bsb and Google-books collection.

2.5 Automatic Document Image Binarization using Bayesian Optimization

In research [5] Ekta Vats, Anders Hast and Prashant Singh proposed an automatic document image binarization algorithm to segment the text from heavily degraded document images to overcome the problem of various forms of degradation and its effect. They proposed a two band-pass filtering approach for the background noise removal and Bayesian optimization for automatic hyper parameter selection for the optimal results.

They have tested their proposed method on the Document Image Binarization competition (DIBCO) and other competitions and got effective results.

Here are the results which shows how much their method improved upon all other related algorithms-

Methods	DIBCO 2009-2016		DIBCO 2009-2013		DIBCO 2011-2014		
	F-measure (%) (")	PSNR (")	F-measure (%) (")	PSNR (")	F-measure (%) (")	PSNR (")	DRD (#)
Otsu [10]	84.10	16.68	82.06	16.05	84.53		16.53
Sauvola [12]	82.93	16.54	82.23	16.35	84.32		16.76
LMM [17]	-	-	89.15	18.78	-		-
Howe [1]	-	-	-	-	91.88		20.83
Proposed method	90.99	19.00	90.21	18.71	91.16		19.09

Table 6: Comparison results of average F-measure (%), PSNR and DRD values obtained using different binarization methods.

2.6 A Shadow Removal Method for Tesseract Text Recognition

In research [6] Huimin Lu, Baofang Guo , Juntao Liu and Xijun Yan proposed a new method to process the shadowed text images for the Tesseract's optical character recognition engine where a local adaptive threshold algorithm is used to transform the grayscale image into a binary image to capture the contours of the text. Then to delete the salt and pepper noise in the shadowed areas they proposed a double filtering algorithm in which a projection method is used to remove the noise between texts and the median filter is used to remove the noise within characters. And then the processed image is fed to the tesseract engine.

In order to show the superiority, they tested their proposed method side by side with AHE method and gamma correction method with the same test dataset.

Here is the result evaluation of the proposed method side by side with the AHE and gamma correction method-

Test Image	No Preprocessing	AHE Method	Gamma Correction	Our Scheme
T1	78.2%	80.3%	81.6%	88.8%
T2	86.5%	86.5%	86.5%	94.3%
T3	86.9%	88.4%	89.6%	94.6%
T4	75.5%	73.2%	77.1%	84.2%
T5	92.0%	92.0%	92.7%	97.8%

Table 7: OCR accuracy comparison

2.7 Insights on the Use of Convolutional Neural Networks for the Document Image Binarization

Convolutional Neural networks have and had significant performance in computer vision problems and Handwritten text recognition problems. In research [7] J. Pastor-Pellier, S. Espana-Boquera , F. Zamora Martinez , M. Zeshan Afzal and Maria Jose Castro-Bleda proposed the use of CNN models to document image binarization. Their main idea was to classify each pixel of the image into foreground and background from a sliding window centered at the pixel to be classified. They did an experimental analysis on DIBCO and Santgall datasets.

They applied a set of convolution and max-pooling transforms to the input image. Convolution is able to learn useful features and several kernels are used to obtain a set of maps.

They have used the data from DIBCO 2009, H-DIBCO 2010, DIBCO 2011, H-DIBCO 2012 for training sets and DIBCO 2013 as the final test set. They have also used noisy images from the Historic IAM Santgall Database.

Comparison of the binarization performance for CNNs and other methods for the test sets of DIBCO and Santgall databases -

Method	DIBCO 2013			Santgall		
	FM	MSE	PSNR	FM	MSE	PSNR
Otsu	83.94	0.056	16.94	80.71	0.020	17.09
Sauvola	85.02	0.047	16.63	88.68	0.010	19.86
MLP	82.31	0.029	16.89	93.94	0.005	22.80
MLP+Features	85.82	0.021	18.18	94.75	0.005	23.39
CNN	87.74	0.020	18.91	97.02	0.002	27.22

Table 8: Comparison of binarization performance for CNNs and other methods for the test sets of DIBOC and Santgall databases

2.8 Can fully Convolutional networks perform well for general image restoration problem?

In research [8] Subhajit Chaudhury and Hiya Roy proposed a fully convolutional network(FCN) based approach for color image restoration. So they proposed a fully convolutional model that learns a direct end-to-end mapping between the degraded images as input and the desired clean images as output. Their model is on domain transformation technique but represent a data-driven task specific approach. Novel basis projection, task dependent coefficient alteration and image restoration are represented as convolutional networks.

Their approach outperforms other traditional denoising methods and shows exceptional results. Their model also done amazing job on blind image reconstruction.

For training their model they used data from 1) ImageNet and 2) MSCOCO and for testing they used 1) Berkeley Segmentation dataset and 2) Pascal VOC 2012.

Image denoising performances for Berkeley segmentation dataset images are -

Image	= 25			= 50	
	KSVD	CBM3D	IRCNN	CBM3D	IRCNN
Castle	31.19	32.24	32.17	28.67	28.66
Mushroom	30.26	31.20	30.92	27.77	27.60
Horse	29.81	30.67	30.83	27.59	27.84
Kangaroo	28.39	29.19	29.30	26.37	26.45
Train	28.16	28.72	28.88	24.52	25.06
Average	29.56	30.40	30.42	26.98	27.12

Table 9: Image denoising performance for Berkeley segmentation dataset images

2.9 Enhancing OCR Accuracy with Super Resolution

Quality of the image often suppresses the accuracy of OCR engines for document images. Performance degradation of this type is mainly happening for poor quality of scanning and poor resolution. In research [9] Ankit Lat and C.V. Jawahar proposed a method for super resolving document images using Generative Adversarial Network(GAN) to solve the above problem.

They proposed a super resolution based preprocessing step that can enhance the accuracies of the OCRs. Their method is tested and suited for printed document images.

They have used 3 types of datasets to test their system. English Novel Dataset, Cross Language Dataset and Scanned Web Dataset.

Their proposed Method showed an improvement up to 21 % in accuracy in OCR on test images scanned at low resolution.

2.10 Cleaning Up Dirty Scanned Documents with Deep Learning

In Article [10] the author explained various cleaning methods of dirty noisy images. CNN Autoencoder and CNN stacker was the main premises of the article. The article explained how CNN Autoencoder and CNN stacking overcome the error of traditional cleaning methods. And both of them achieved very low loss rate on the testing.

Autoencoders are neural network composed of an encoder and a decoder. The encoder compresses the input data into a lower dimensional representation The decoder reconstructs the

representation to obtain an output that mimics the input as closely as possible. During this process autoencoder learns the silent features of the input data.

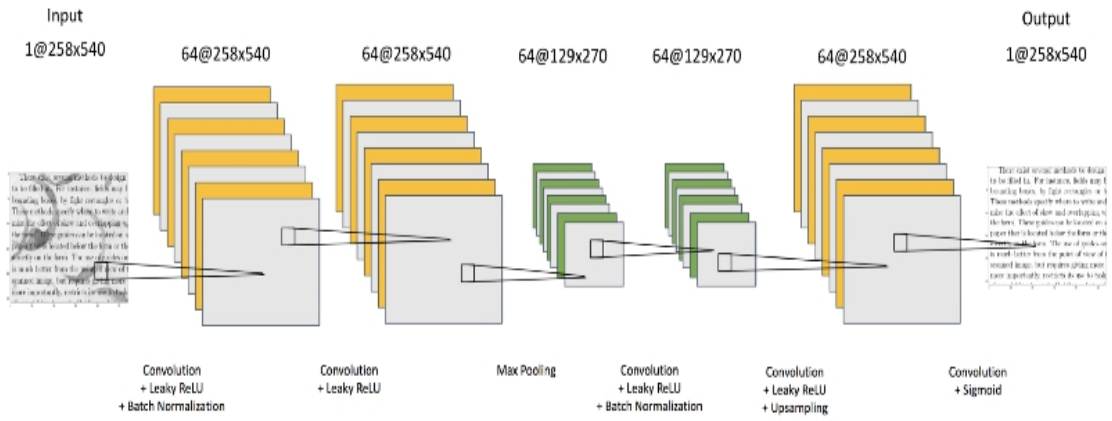


Figure 7: CNN Autoencoder architecture [10]

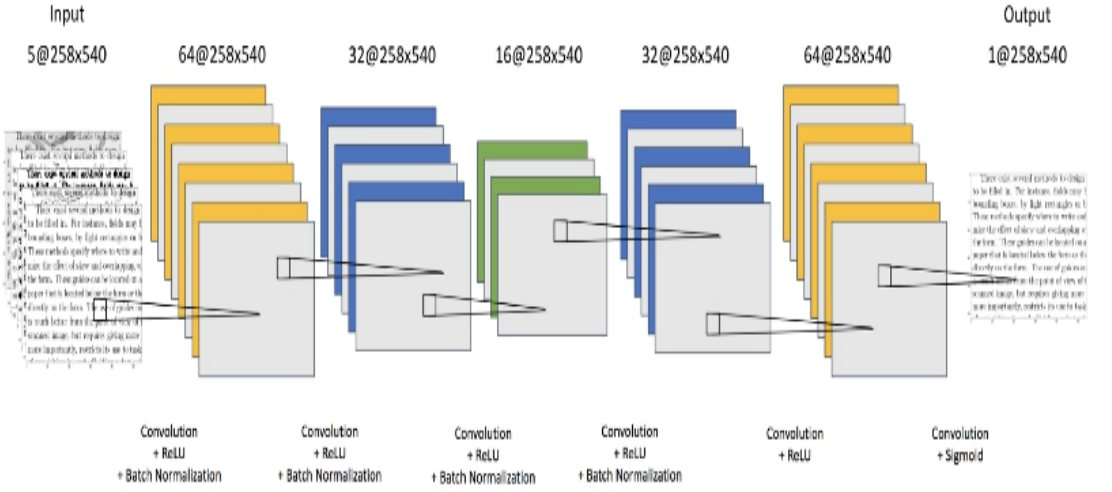


Figure 8: CNN stacker architecture [10]

The CNN Autoencoder alone achieves RMSE (Root Mean Squire Error) of 0.025 on testing The CNN stacker model bring down the RMSE to 0.019 on testing.

2.11 Summary of the Literature

For this research we have studied many research paper related to our research work. Among them we have selected 10 articles. From background we have observed that many of the articles provided a solution for degrading document image preservation. But doing so they all didn't mention about quality of the output images and how it effects the result of the OCR. Some of the articles went with the traditional approaches and other used some modern approaches. Traditional approaches are lagging behind than modern approaches. We think that modern deep neural network approaches are more efficient and time saving approach for image pre-processing. We tried to go further into the concept of using modern deep neural network approaches for pre-processing and getting even better OCR performance.

Chapter 3

METHODOLOGY

3.1 Collection of Dataset

For generating good outputs first target should be good set of datasets. So, we have collected 2 separate datasets:

1) DIBCO (2009-2018) [11]

2) Kaggle Dataset (2015) [12]

We also have some extra manually created datasets [13]. First we collected different sized images from some novels and we have created total 550 noisy images using various kind of image noises. We separated these data into several parts: training testing and validation.

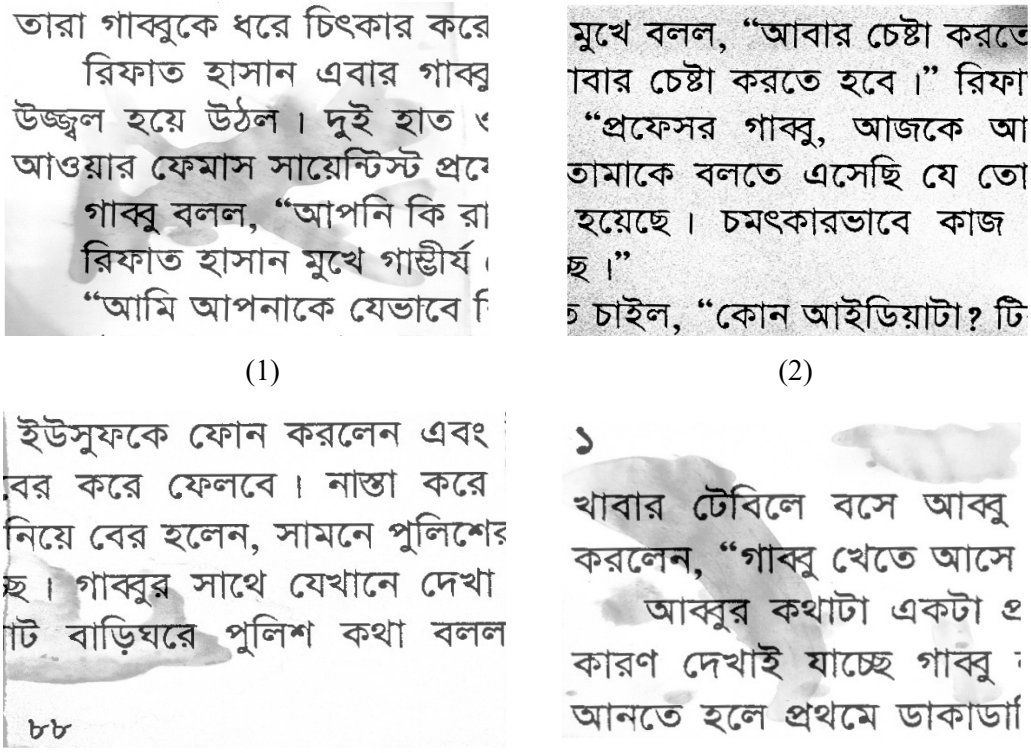


Figure 9: Our created dataset (1-4) [13]

3.2 Previous work done

We implemented several paper and article [10] method implementation and coded the process of cleaning dirty images. We trained and tested the models and got the expected outputs that was told on the paper.

3.2.1 Challenges

We have studied and implemented some of the papers and got the desired output but faced some challenges along the way. Challenges being-

1) **Challenge 1:** When implementing and testing the article [10] we saw that we need to scale the images to a specific format so the model can be trained and tested. But this converts high resolution images into low resolution which later affect the output and it didn't give expected outcome on OCR engine.

2) **Challenge 2:** On the other hand, when testing any poor quality image, we have excessive amount of information loss that we cannot have proper OCR output as well

3) Challenge 3: Another problem was that traditional methods are more suitable for handling certain types of noises at a time. So different noises are not handled properly through all the papers.

3.3 Methodology

Our research is based on improving the pre-processing method for better OCR performance. Though its being explored many times, there are un-reached territory and one of them is being the consistency of quality in document images. So our intention was to create a method that tackles the above challenges and whatever the image quality being set we need to pre-process it without any loss of data for OCR performance.

We have used different techniques to tackle different problems. We have used:

1) Zero Padding for small images
2) Cropping and reconstruction of image for large images.

We trained our CNN-Autoencoder model with fixed scaled images of 540*420. Because We tried different combination of image scaling with zero padding and cropping. But if we tried to scale up like- train very large images, then we had to add excessive amount of padding to lower scale images. Which later affect our cleaning process and result is not up to the mark. On the other hand, when we tried a lower scale than we had to crop our images into very small parts and for only one image we had excessive amount of cropped images, which made the model training process a bit complex. We had a vast variety of ranged images. So we had to readjust the scale and fine tune our model results. And we saw that near 540*420 scale our model started to produce better results.

We used a 5 layer CNN-Autoencoder model. The used model is given below with an example:

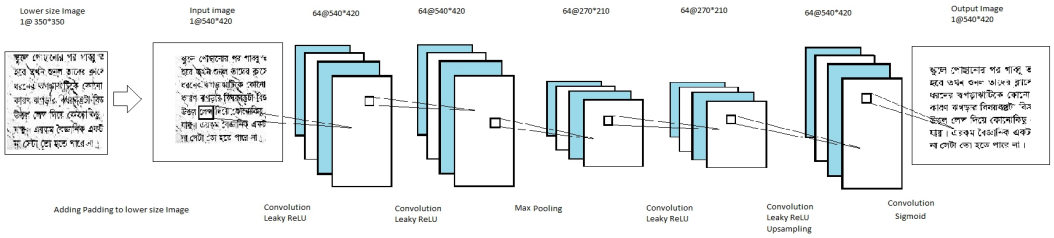


Figure 10: Our CNN Autoencoder Model

Our model's first 2 layers are 2 convolutional layers and max pooling is used after that. There are 3 more convolutional layers. The first of which is a single convolution layer, second layer is another convolution layer with upsampling. And lastly we used a single convolution layer which is connected with the output layer. Noise pattern learning and reconstruction of the original image is done in these last 3 layers. We used Leaky ReLU in hidden layers for handling gradient vanishing problem in hidden layers and sigmoid activation function on the last layer is used for getting the same output range as input.

For testing, our input taking system is a bit different than usual. First we check whether it is smaller or larger sized image and then we added padding or crop the image accordingly to feed it into our model.

Here is the working mechanism of our designed system for our thesis:

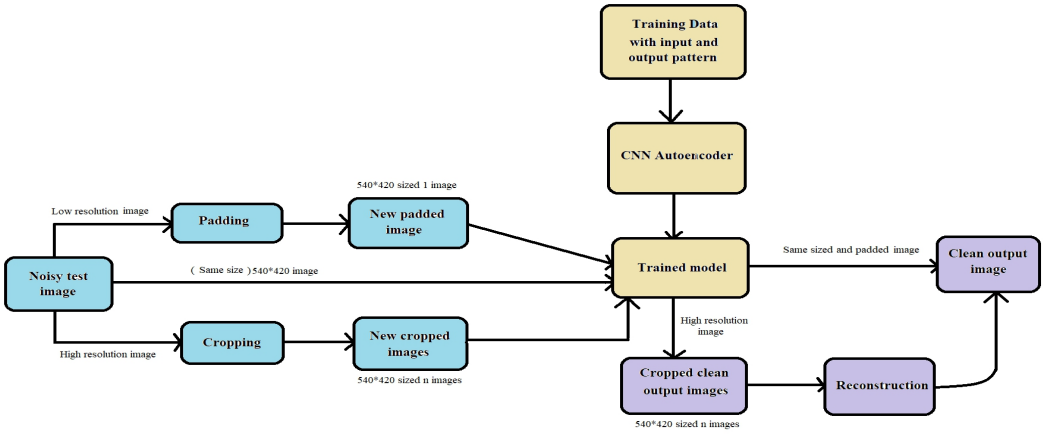


Figure 11: Our System Architecture

Below we will describe each of the step we took in order to solve our challenges:

3.3.1 Adding Zero Padding to small images

When a low resolution image smaller than our trained image size is given as input our system automatically add padding to that image so that it can be as the same size as our desired size of 540*420. We used Zero padding for adding padding to the images so that no addition information is added, just the size of the image increases to our desired level. And our model removes the noise from that image.

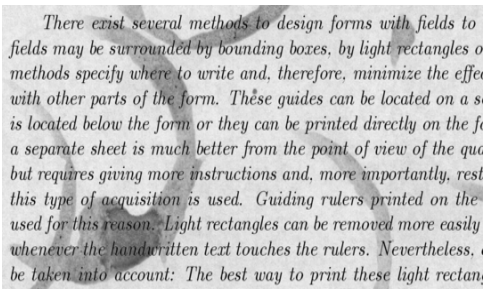


Figure 12: Low resolution noisy image

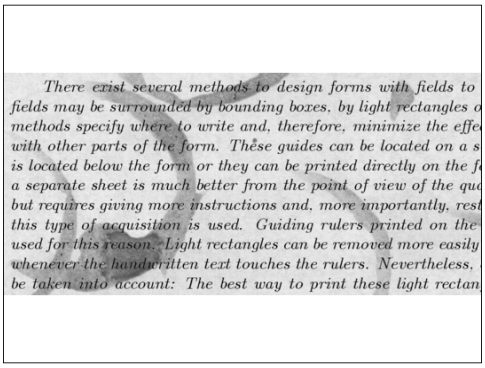


Figure 13: Noisy padded image

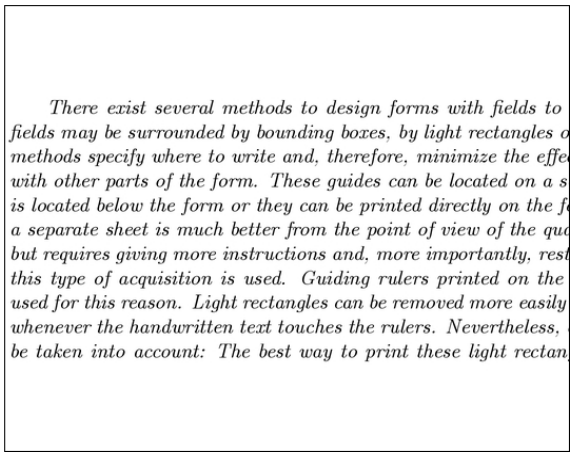


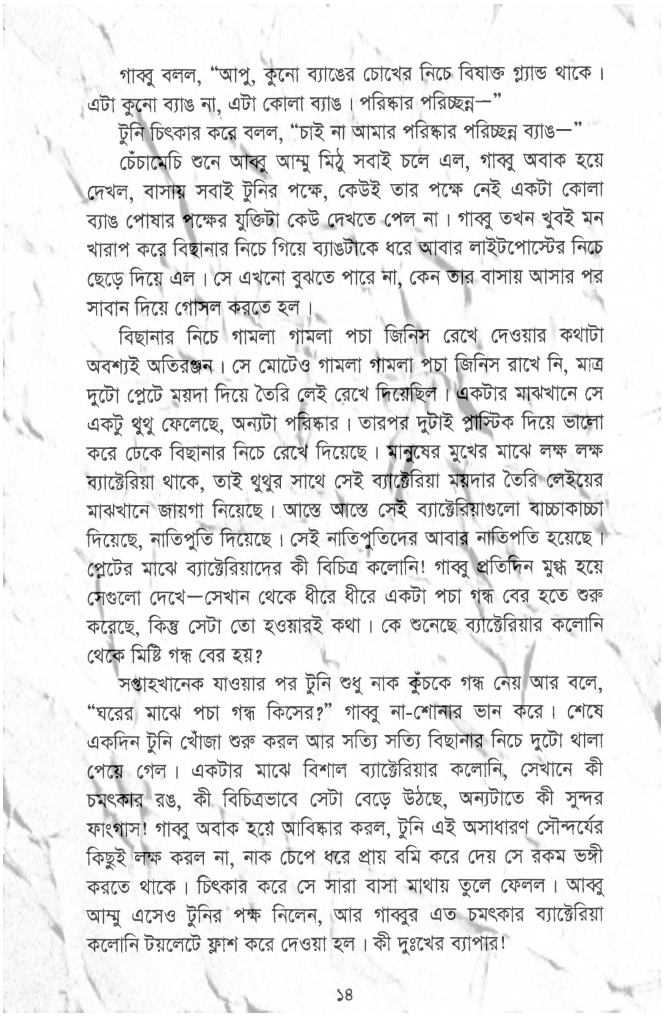
Figure 14: Clean image

3.3.2 Handling different noises at one time

Handling different noises was done successfully using CNN-Autoencoder. Because using deep learning we trained our model to learn all the noise pattern of the collected datasets and recognize them later when testing part is done. So we trained our model with different noise images and our model learned all the noise pattern and gave us the expected output for all of them.

3.3.3 Cropping and reconstruction of image for large images

Large scale images had different problem. So when a large image is given as input previously we needed to scale it down in order to clean it and thus the results were not up to the mark. Now we found a solution to tackle that problem. Right before going into our model we just crop it into several part and make all the part size same (padding is used if there is any uneven size concern). Then we pass these parts one at a time into our modal and clean them, later we just reconstruct the image by reconnecting all the clean separated parts into one image.



গাব্বু বলল, "আপু, কুনো ব্যাঙের চোখের নিচে বিষাক্ত গ্ল্যান্ড থাকে। এটা কুনো ব্যাঙ না, এটা কোলা ব্যাঙ। পরিষ্কার পরিচ্ছন্ন—"

টুনি চিৎকার করে বলল, "চাই না আমার পরিষ্কার পরিচ্ছন্ন ব্যাঙ—"

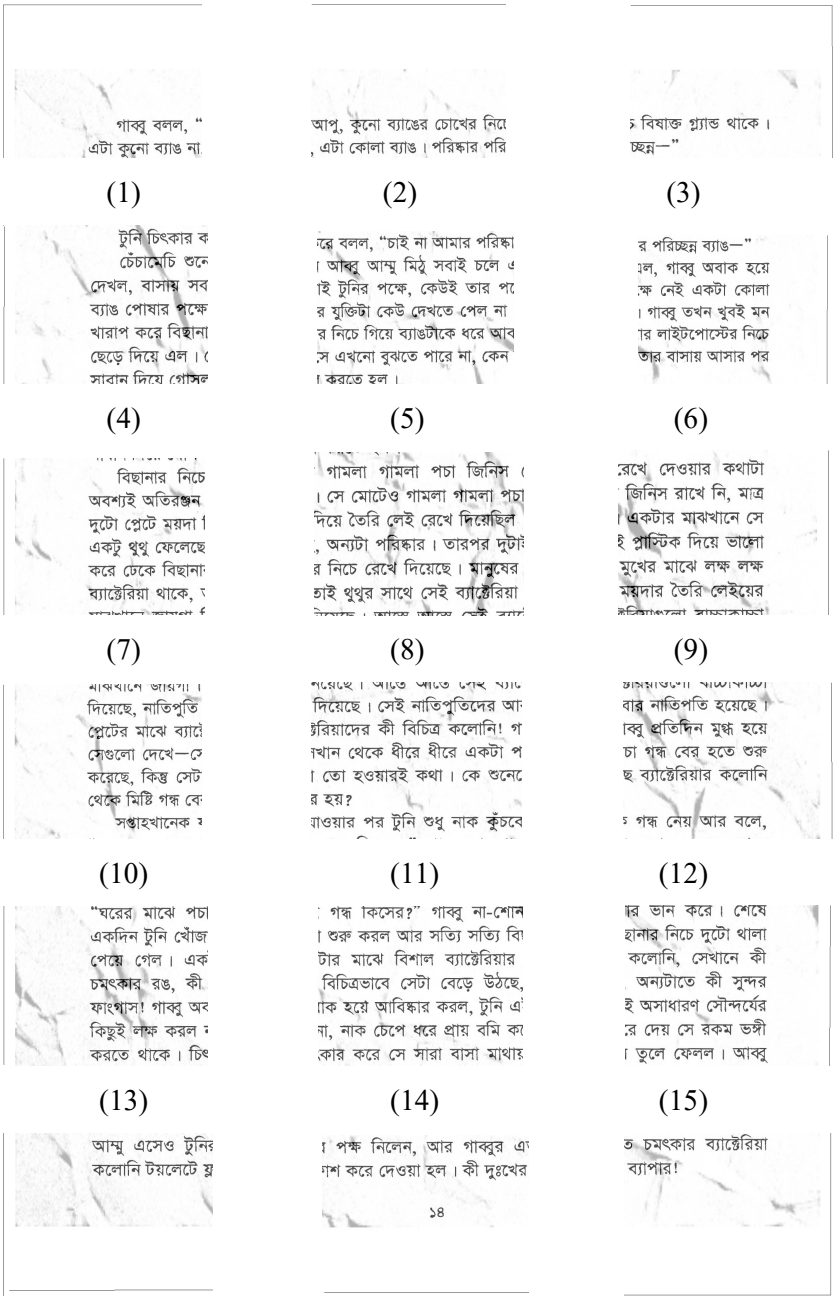
চেঁচামেচি শুনে আব্বু আম্মু মিঠু সবাই চলে এল, গাব্বু অবাক হয়ে দেখল, বাসায় সবাই টুনির পক্ষে, কেউই তার পক্ষে নেই একটা কোলা ব্যাঙ পোষার পক্ষের যুক্তিটা কেউ দেখতে পেল না। গাব্বু তখন খুবই মন খারাপ করে বিছানার নিচে গিয়ে ব্যাঙটাকে ধরে আবার লাইটপোস্টের নিচে ছেড়ে দিয়ে এল। সে এখনো বুঝতে পারে না, কেন তার বাসায় আসার পর সাবান দিয়ে গোসল করতে হল।

বিছানার নিচে গামলা গামলা পচা জিনিস রেখে দেওয়ার কথাটা অবশ্যই অতিরঞ্জন। সে মোটেও গামলা গামলা পচা জিনিস রাখে নি, মাত্র দুটো প্লেটে ময়দা দিয়ে তৈরি লেই রেখে দিয়েছিল। একটার মাঝখানে সে একটু থুথু ফেলেছে, অন্যটা পরিষ্কার। তারপর দুটাই প্লাস্টিক দিয়ে ভালো করে ঢেকে বিছানার নিচে রেখে দিয়েছে। মানুষের মুখের মাঝে লক্ষ লক্ষ ব্যাক্টেরিয়া থাকে, তাই থুথুর সাথে সেই ব্যাক্টেরিয়া ময়দার তৈরি লেইয়ের মাঝখানে জায়গা নিয়েছে। আস্তে আস্তে সেই ব্যাক্টেরিয়াগুলো বাচ্চাকাচ্চা দিয়েছে, নাতিপুতি দিয়েছে। সেই নাতিপুতিদের আবার নাতিপতি হয়েছে। প্লেটের মাঝে ব্যাক্টেরিয়াদের কী বিচিত্র কলোনি! গাব্বু প্রতিদিন মুগ্ধ হয়ে সেগুলো দেখে—সেখান থেকে ধীরে ধীরে একটা পচা গন্ধ বের হতে শুরু করেছে, কিন্তু সেটা তো হওয়ারই কথা। কে শুনেছে ব্যাক্টেরিয়ার কলোনি থেকে মিষ্টি গন্ধ বের হয়?

সপ্তাহখানেক যাওয়ার পর টুনি শুধু নাক কুঁচকে গন্ধ নেয় আর বলে, "ঘরের মাঝে পচা গন্ধ কিসের?" গাব্বু না-শোনার ভান করে। শেষে একদিন টুনি খোঁজা শুরু করল আর সত্যি সত্যি বিছানার নিচে দুটো থালা পেয়ে গেল। একটার মাঝে বিশাল ব্যাক্টেরিয়ার কলোনি, সেখানে কী চমৎকার রঙ, কী বিচিত্রভাবে সেটা বেড়ে উঠছে, অন্যটাতে কী সুন্দর ফাংগাস! গাব্বু অবাক হয়ে আবিষ্কার করল, টুনি এই অসাধারণ সৌন্দর্যের কিছুই লক্ষ করল না, নাক চেপে ধরে প্রায় বমি করে দেয় সে রকম ভঙ্গী করতে থাকে। চিৎকার করে সে সারা বাসা মাথায় তুলে ফেলল। আব্বু আম্মু এসেও টুনির পক্ষ নিলেন, আর গাব্বুর এত চমৎকার ব্যাক্টেরিয়া কলোনি টয়লেটে ফ্লাশ করে দেওয়া হল। কী দুঃখের ব্যাপার!

১৪

Figure 15: High resolution noisy image



(1) (2) (3)

(4) (5) (6)

(7) (8) (9)

(10) (11) (12)

(13) (14) (15)

(16) (17) (18)

Figure 16: Cropped images (1-18)

গাব্বু বলল, "আপু, কুনো ব্যাঙের চোখের নিচে বিষাক্ত গ্ল্যান্ড থাকে। এটা কুনো ব্যাঙ না, এটা কোলা ব্যাঙ। পরিষ্কার পরিচ্ছন্ন—"

টুনি চিৎকার করে বলল, "চাই না আমার পরিষ্কার পরিচ্ছন্ন ব্যাঙ—"

চেঁচামেচি শুনে আব্বু আম্মু মিঠু সবাই চলে এল, গাব্বু অবাক হয়ে দেখল, বাসায় সবাই টুনির পক্ষে, কেউই তার পক্ষে নেই একটা কোলা ব্যাঙ পোষার পক্ষের যুক্তিটা কেউ দেখতে পেল না। গাব্বু তখন খুবই মন খারাপ করে বিছানার নিচে গিয়ে ব্যাঙটাকে ধরে আবার লাইটপোস্টের নিচে ছেড়ে দিয়ে এল। সে এখনো বুঝতে পারে না, কেন তার বাসায় আসার পর সাবান দিয়ে গোসল করতে হল।

বিছানার নিচে গামলা গামলা পচা জিনিস রেখে দেওয়ার কথাটা অবশ্যই অতিরঞ্জন। সে মোটেও গামলা গামলা পচা জিনিস রাখে নি, মাত্র দুটো প্লেটে ময়দা দিয়ে তৈরি লেই রেখে দিয়েছিল। একটার মাঝখানে সে একটু থুথু ফেলেছে, অন্যটা পরিষ্কার। তারপর দুটাই প্লাস্টিক দিয়ে ভালো করে ঢেকে বিছানার নিচে রেখে দিয়েছে। মানুষের মুখের মাঝে লক্ষ লক্ষ ব্যাক্টেরিয়া থাকে, তাই থুথুর সাথে সেই ব্যাক্টেরিয়া ময়দার তৈরি লেইয়ের মাঝখানে জায়গা নিয়েছে। আস্তে আস্তে সেই ব্যাক্টেরিয়াগুলো বাচ্চাকাচ্চা দিয়েছে, নাতিপুতি দিয়েছে। সেই নাতিপুতিদের আবার নাতিপতি হয়েছে। প্লেটের মাঝে ব্যাক্টেরিয়াদের কী বিচিত্র কলোনি! গাব্বু প্রতিদিন মুগ্ধ হয়ে সেগুলো দেখে—সেখান থেকে ধীরে ধীরে একটা পচা গন্ধ বের হতে শুরু করেছে, কিন্তু সেটা তো হওয়ারই কথা। কে শুনেছে ব্যাক্টেরিয়ার কলোনি থেকে মিষ্টি গন্ধ বের হয়?

সপ্তাহখানেক যাওয়ার পর টুনি শুধু নাক কুঁচকে গন্ধ নেয় আর বলে, "ঘরের মাঝে পচা গন্ধ কিসের?" গাব্বু না-শোনার ভান করে। শেষে একদিন টুনি খোঁজা শুরু করল আর সত্যি সত্যি বিছানার নিচে দুটো থালা পেয়ে গেল। একটার মাঝে বিশাল ব্যাক্টেরিয়ার কলোনি, সেখানে কী চমৎকার রঙ, কী বিচিত্রভাবে সেটা বেড়ে উঠছে, অন্যটাতে কী সুন্দর ফাংগাস! গাব্বু অবাক হয়ে আবিষ্কার করল, টুনি এই অসাধারণ সৌন্দর্যের কিছুই লক্ষ করল না, নাক চেপে ধরে প্রায় বমি করে দেয় সে রকম ভঙ্গী করতে থাকে। চিৎকার করে সে সারা বাসা মাথায় তুলে ফেলল। আব্বু আম্মু এসেও টুনির পক্ষ নিলেন, আর গাব্বুর এত চমৎকার ব্যাক্টেরিয়া কলোনি টয়লেটে ফ্লাশ করে দেওয়া হল। কী দুঃখের ব্যাপার!

১৪

Figure 17: High resolution clean image

Chapter 4

RESULTS AND DISCUSSION

We have collected different datasets and trained our model with those datasets and tested out different noisy images. Though previous systems gave us different results due to variable size images and different noise category. We worked with 2 target in our mind. One is being handling the variable size image and another one is handling different noises.

Results for those parts are given below:

4.1 Pre-processing Results

In previous system, for high and low resolution, images are either been scaled down or scaled up. which distorted the images and produce a blurry or stretched out image output. So we had to use a particular sized image.

4.1.1 Our Previous work

বিশ্ব ডায়াবেটিস দিবস আজ। বিশ্ব জুড়ে ডায়াবেটিস সম্পর্কে সচেতনতা বৃদ্ধির লক্ষ্যে প্রতি বছর আজকের এই দিনে পালন করা হয় বিশ্ব ডায়াবেটিস দিবস। সারা বিশ্বে ডায়াবেটিস রোগ ব্যাপক হারে বেড়ে যাওয়ায় বিশ্ব ডায়াবেটিস ফেডারেশন ও বিশ্ব স্বাস্থ্য সংস্থা ১৯৯১ সালে ১৪ই নভেম্বরকে বিশ্ব ডায়াবেটিস দিবস ঘোষণা করে। উদ্যোগ করা হয়েছে পরে, ফিরদৌসী ফেডরিক বেনটিং এর জন্ম দিন এই ১৪ই নভেম্বরকেই বিশ্ব ডায়াবেটিস দিবস হিসেবে ঘোষণা করা হয়। এই ফেডরিক বেনটিংই ডায়াবেটিস রোগের চিকিৎসায় ব্যবহৃত ইনসুলিন আবিষ্কার করেন। অবশ্য এই ইনসুলিন আবিষ্কারের সঙ্গে আরেকজন বিজ্ঞানী চার্লস বেস্টও সম্পৃক্ত ছিলেন। ডায়াবেটিসকে মানুষের জন্য একটা 'নীরব ঘাতক' হিসেবেই অভিহিত করেছেন বিজ্ঞানীগণ। আর বাংলাদেশে নেহেতু ডায়াবেটিসে আক্রান্তের সংখ্যা দিন দিন বেড়ে চলেছে, তাই এখানে দিবসটি পালনের গুরুত্ব অপরিসীম।

এখানে ডায়াবেটিসে আক্রান্তের সংখ্যা দিন দিন বেড়ে চলেছে। নারী-পুরুষ-শিশু সব বয়সী মানুষই এখন ডায়াবেটিসে আক্রান্ত হচ্ছে। শহরাঞ্চলের সঙ্গে পাল্লা দিয়ে গ্রামাঞ্চলেও এই ঘাতক ব্যাধি ছড়িয়ে পড়ছে। পরিসংখ্যানের তথ্য হচ্ছে দেশে বর্তমানে ৮৪ লাখের বেশি লোক ডায়াবেটিসে আক্রান্ত। এর মধ্যে ৪৫ লাখ রোগীকে চিকিৎসা সেবার আওতায় নিয়ে আসা সম্ভব হচ্ছে না। জরিপে বেরিয়ে আসা আরেকটি ভয়াবহ তথ্য হচ্ছে, আক্রান্ত হওয়ে প্রায় ৫০ শতাংশে মানুষই জানেন না যে তিনি ডায়াবেটিস বহন করছেন। এর মূল কারণ সচেতনতার অভাব। কারণ বড় ধরনের অসুস্থতা ছাড়া এদেশের বেশির ভাগ মানুষই শারীরিক পরীক্ষা-নিরীক্ষায় অভ্যস্থ নন। বিশ্বের অনেক দেশ ডায়াবেটিস নিয়ন্ত্রণে যথেষ্ট উন্নতি এবং বিভিন্ন মাত্রায় সফল। কিন্তু বাংলাদেশের চিত্র ভিন্ন। ডায়াবেটিসের দীর্ঘকালীন জটিলতায় বাংলাদেশের রোগীরা বেশি ভুগছে। সবচেয়ে ভয়াবহ হচ্ছে, অতি অল্প বয়সীরা টাইপ-২ ডায়াবেটিসে আক্রান্ত হচ্ছে। বর্তমানে ডায়াবেটিস রোগীর সংখ্যা বিবেচনায় বিশ্বে বাংলাদেশের অবস্থান দশম স্থানে। রোগী বৃদ্ধির বর্তমান হার অব্যাহত থাকলে ২০৩৩ সালে ডায়াবেটিস রোগীর সংখ্যা বর্তমানের চেয়ে দ্বিগুণ হবে বলে আশংকা করেছে।

ডায়াবেটিস একটি বিপাকজনিত রোগ। মানবদেহে ইনসুলিন নামক হরমোনের ঘাটতি হলে কিংবা উৎপাদিত ইনসুলিন কার্যকরভাবে শরীরে ব্যবহৃত না হলে বা শরীরের ইনসুলিন নিষ্ক্রিয়তা থাকলে রক্তে গ্লুকোজের পরিমাণ স্বাভাবিকের চেয়ে বেড়ে যায়। যা প্রস্রাবের সঙ্গে বেরিয়ে যায়। এই অবস্থার নামই ডায়াবেটিস। এটি কোন সংক্রামক রোগ নয়, তবে এটি সংক্রামক রোগের মতোই দ্রুত ছড়িয়ে পড়ছে। আর সচেতনতার মাধ্যমে তথা খাদ্যাভ্যাস ও জীবনাচারে একটা সুশৃঙ্খল নিয়ম মেনে চললে ডায়াবেটিসকে হারে রাখা সম্ভব। পাশাপাশি সব সরকারি-বেসরকারি হাসপাতালে এ রোগের চিকিৎসা যাতে রোগীরা সহজে পায়, সে ব্যবস্থা করতে হবে।

Figure 18: High resolution noisy image

Figure 19: Scaled down clean image

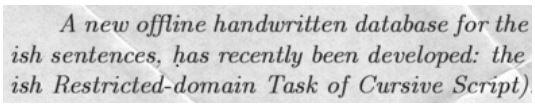


Figure 20: low resolution noisy image

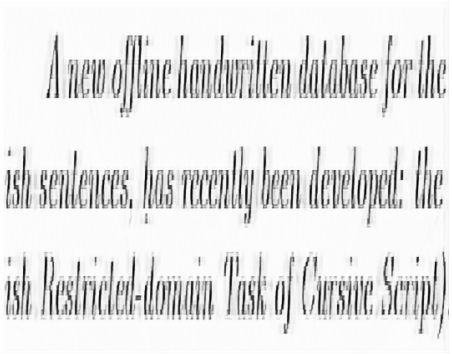


Figure 21: scaled up clean image

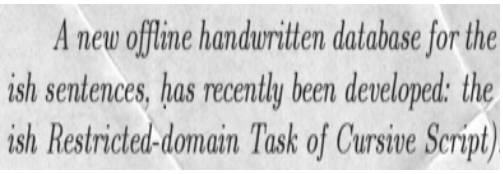
4.1.2 Our Final Work

We worked with two different techniques. And the results are significantly better.

In our system, for variable sized image we used zero padding for lower resolution image and cropping and reconstruction technique for high resolution. Which produced a much better output for variable sized images.

Figure 22: High resolution noisy image

Figure 23: Our system cleaned image



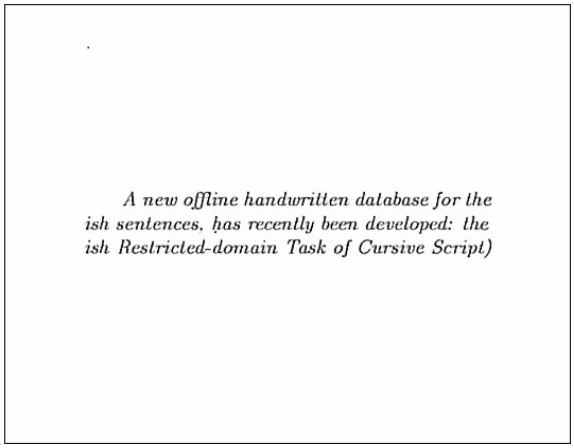


Figure 24: Low resolution noisy image

Figure 25: Our system padded clean image

4.2 OCR Test Results

Pre-processing is very necessary step for OCR testing. Without preprocessing the OCR accuracy is much lower than with processing.

We tested different scenario to prove that and also show that how much improvement can be done using our method

4.2.1 Without pre-processing OCR results

Without any preprocessing OCR output is very poor.

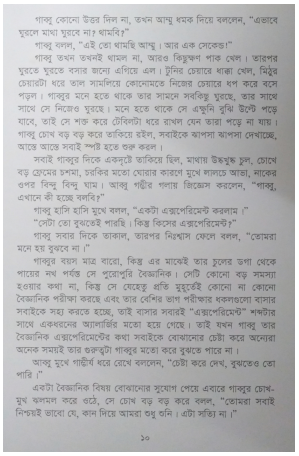


Figure 26: Noisy image

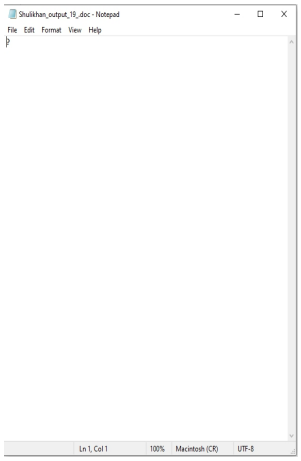


Figure 27: OCR for Shulikhan

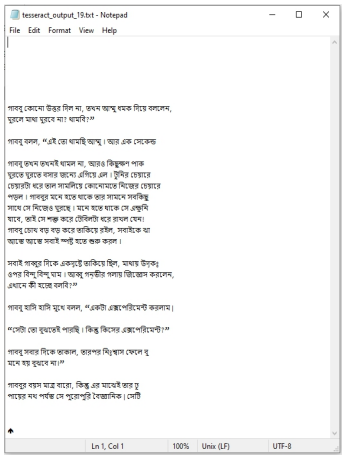
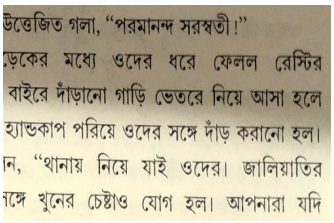
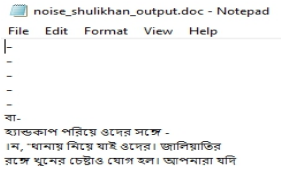


Figure 28: OCR for Tesseract





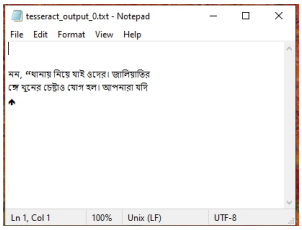


Figure 29: Noisy image **Figure 30: OCR for Shulikhan** **Figure 31: OCR for Tesseract**

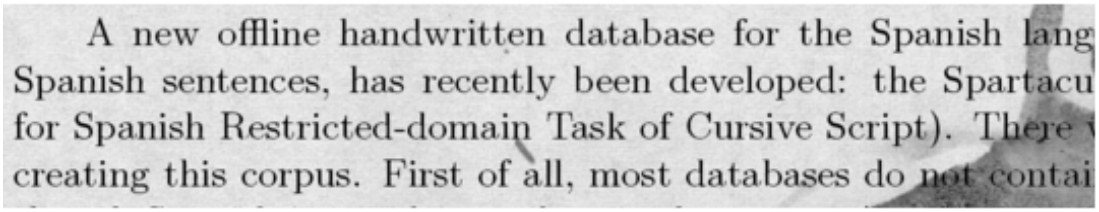


Figure 32: Noisy image

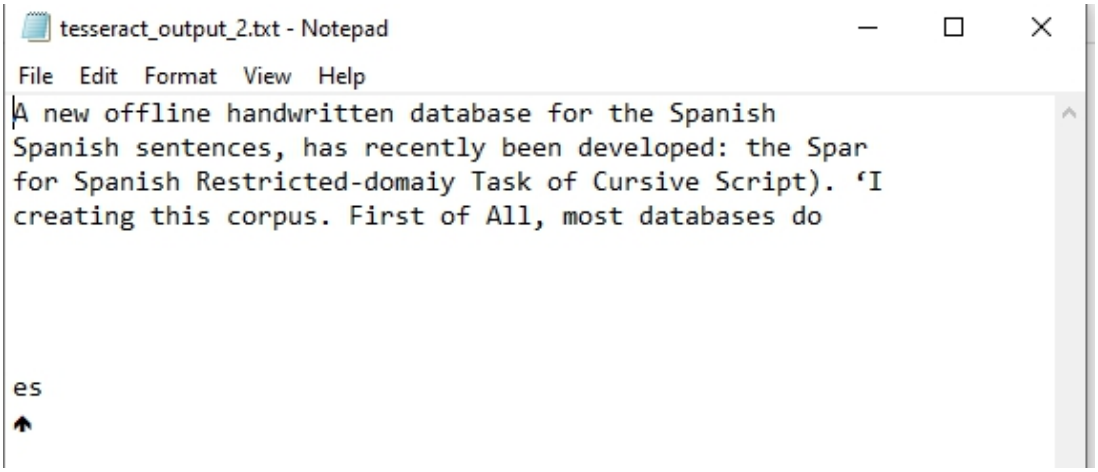


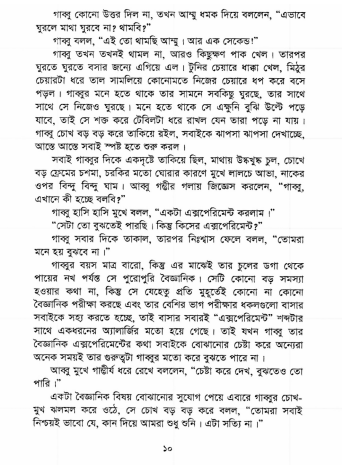
Figure 33: Tesseract output

4.2.2 Previous work OCR results

There are various denoising techniques exits. But none of them handles variable size image denoising. So for high and low resolution image cleaned images are more distorted, hence the OCR output is too low and not understandable.

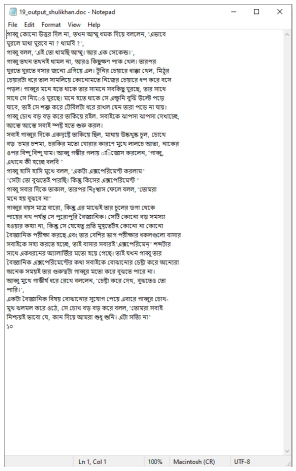
4.2.3 Final work OCR Results

In our system, we handled all the issues that we were having. We handled variable size image cleaning using padding and cropping-reconstruction techniques and then cleaned the images in our CNN Autoencoder network model. So the output that our system produces are much cleaner and OCR results are significantly better.



গাব্বু কোনো উত্তর দিল না, তখন আব্বু ধমক দিয়ে বললেন, "এভাবে
ঘুরলে মাথা ঘুরবে না? থামবি?"
গাব্বু বলল, "এই তো থামছি আব্বু। আর এক সেকেন্ড।"
গাব্বু তখন তখনই থামল না, আরও কিছুক্ষণ পাক খেল। তারপর
ঘুরতে ঘুরতে বসার জন্যে এগিয়ে এল। টুলিন চেয়ারে ধাক্কা খেল, মিঠুর
চেয়ারটা ধরে তাল সামলিয়ে কোনোমতে নিজের চেয়ারে ধপ করে বসে
পড়ল। গাব্বুর মনে হতে থাকে তার সামনে সবকিছু ঘুরছে, তার সাথে
সাথে সে নিজেও ঘুরছে। মনে হতে থাকে সে এক্ষুনি বুঝি উল্টে পড়ে
যাবে, তাই সে শক্ত করে টেবিলটা ধরে রাখল যেন তারা পড়ে না যায়।
গাব্বু চোখ বড় বড় করে তাকিয়ে রইল, সবাইকে কাপসা কাপসা দেখাচ্ছে,
আস্তে আস্তে সবাই স্পষ্ট হতে শুরু করল।
সবাই গাব্বুর দিকে একদৃষ্টে তাকিয়ে ছিল, মাথায় উস্কখুস্ক চুল, চোখে
বড় ফ্রেমের চশমা, চরকির মতো ঘোরার কারণে মুখে লালচে আভা, নাকের
ওপর বিন্দু বিন্দু ঘাম। আব্বু গম্ভীর গলায় জিজ্ঞেস করলেন, "গাব্বু,
এখানে কী হচ্ছে বলবি?"
গাব্বু হাসি হাসি মুখে বলল, "একটা এক্সপেরিমেন্ট করলাম।"
"সেটা তো বুঝতেই পারছি। কিন্তু কিসের এক্সপেরিমেন্ট?"
গাব্বু সবার দিকে তাকাল, তারপর নিশ্বাস ফেলে বলল, "তোমরা
মনে হয় বুঝবে না।"
গাব্বুর বয়স মাত্র বারো, কিন্তু এর মাঝেই তার চুলের ডগা থেকে
পায়ের নখ পর্যন্ত সে পুরোপুরি বৈজ্ঞানিক। সেটি কোনো বড় সমস্যা
হওয়ার কথা না, কিন্তু সে যেহেতু প্রতি মুহূর্তেই কোনো না কোনো
বৈজ্ঞানিক পরীক্ষা করছে এবং তার বেশির ভাগ পরীক্ষার ধকলগুলো বাসার
সবাইকে সহ্য করতে হচ্ছে, তাই বাসার সবারই "এক্সপেরিমেন্ট" শব্দটার
সাথে একধরনের অ্যালার্জির মতো হয়ে গেছে। তাই যখন গাব্বু তার
বৈজ্ঞানিক এক্সপেরিমেন্টের কথা সবাইকে বোঝানোর চেষ্টা করে অন্যেরা
অনেক সময়ই তার গুরুত্বটুকু গাব্বুর মতো করে বুঝতে পারে না।
আব্বু মুখে গাম্ভীর্য ধরে রেখে বললেন, "চেষ্টা করে দেখ, বুঝতেও তো
পারি।"
একটা বৈজ্ঞানিক বিষয় বোঝানোর সুযোগ পেয়ে এবার গাব্বুর চোখ-
মুখ ঝলমল করে ওঠে, সে চোখ বড় বড় করে বলল, "তোমরা সবাই
নিশ্চয়ই ভাবো যে, কান দিয়ে আমরা শুধু শুনি। এটা সত্যি না।"

১০

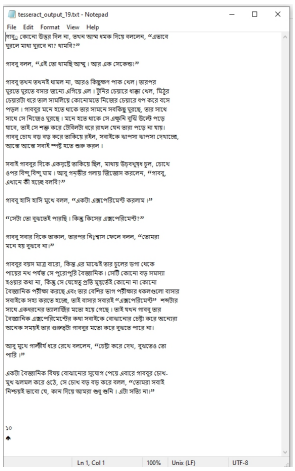


10_output_shulikhan.doc - Notepad

File Edit Format View Help

গাব্বু কোনো উত্তর দিল না, তখন আব্বু ধমক দিয়ে বললেন, 'এভাবে
ঘুরলে মাথা ঘুরবে না? থামবি?'
গাব্বু বলল, 'এই তো থামছি আব্বু। আর এক সেকেন্ড।'
গাব্বু তখন তখনই থামল না, আরও কিছুক্ষণ পাক খেল। তারপর
ঘুরতে ঘুরতে বসার জন্যে এগিয়ে এল। টুলিন চেয়ারে ধাক্কা খেল, মিঠুর
চেয়ারটা ধরে তাল সামলিয়ে কোনোমতে নিজের চেয়ারে ধপ করে বসে
পড়ল। গাব্বুর মনে হতে থাকে তার সামনে সবকিছু ঘুরছে, তার সাথে
সাথে সে নিজেও ঘুরছে। মনে হতে থাকে সে এক্ষুনি বুঝি উল্টে পড়ে
যাবে, তাই সে শক্ত করে টেবিলটা ধরে রাখল যেন তারা পড়ে না যায়।
গাব্বু চোখ বড় বড় করে তাকিয়ে রইল, সবাইকে কাপসা কাপসা দেখাচ্ছে,
আস্তে আস্তে সবাই স্পষ্ট হতে শুরু করল।
সবাই গাব্বুর দিকে একদৃষ্টে তাকিয়ে ছিল, মাথায় উস্কখুস্ক চুল, চোখে
বড় ফ্রেমের চশমা, চরকির মতো ঘোরার কারণে মুখে লালচে আভা, নাকের
ওপর বিন্দু বিন্দু ঘাম। আব্বু গম্ভীর গলায় জিজ্ঞেস করলেন, 'গাব্বু,
এখানে কী হচ্ছে বলবি?'
গাব্বু হাসি হাসি মুখে বলল, 'একটা এক্সপেরিমেন্ট করলাম।'
'সেটা তো বুঝতেই পারছি। কিন্তু কিসের এক্সপেরিমেন্ট?'
গাব্বু সবার দিকে তাকাল, তারপর নিশ্বাস ফেলে বলল, 'তোমরা
মনে হয় বুঝবে না।'
গাব্বুর বয়স মাত্র বারো, কিন্তু এর মাঝেই তার চুলের ডগা থেকে
পায়ের নখ পর্যন্ত সে পুরোপুরি বৈজ্ঞানিক। সেটি কোনো বড় সমস্যা
হওয়ার কথা না, কিন্তু সে যেহেতু প্রতি মুহূর্তেই কোনো না কোনো
বৈজ্ঞানিক পরীক্ষা করছে, এবং তার বেশির ভাগ পরীক্ষার ধকলগুলো বাসার
সবাইকে সহ্য করতে হচ্ছে, তাই বাসার সবারই 'এক্সপেরিমেন্ট' শব্দটার
সাথে একধরনের অ্যালার্জির মতো হয়ে গেছে। তাই যখন গাব্বু তার
বৈজ্ঞানিক এক্সপেরিমেন্টের কথা সবাইকে বোঝানোর চেষ্টা করে অন্যেরা
অনেক সময়ই তার গুরুত্বটুকু গাব্বুর মতো করে বুঝতে পারে না।
আব্বু মুখে গাম্ভীর্য ধরে রেখে বললেন, 'চেষ্টা করে দেখ, বুঝতেও তো
পারি।'
একটা বৈজ্ঞানিক বিষয় বোঝানোর সুযোগ পেয়ে এবার গাব্বুর চোখ-
মুখ ঝলমল করে ওঠে, সে চোখ বড় বড় করে বলল, 'তোমরা সবাই
নিশ্চয়ই ভাবো যে, কান দিয়ে আমরা শুধু শুনি। এটা সত্যি না।'

১০



tesseract_output_19.txt - Notepad

File Edit Format View Help

গাব্বু কোনো উত্তর দিল না, তখন আব্বু ধমক দিয়ে বললেন, "এভাবে
ঘুরলে মাথা ঘুরবে না? থামবি?"

গাব্বু বলল, "এই তো থামছি আব্বু। আর এক সেকেন্ড।"

গাব্বু তখন তখনই থামল না, আরও কিছুক্ষণ পাক খেল। তারপর
ঘুরতে ঘুরতে বসার জন্যে এগিয়ে এল। টুলিন চেয়ারে ধাক্কা খেল, মিঠুর
চেয়ারটা ধরে তাল সামলিয়ে কোনোমতে নিজের চেয়ারে ধপ করে বসে
পড়ল। গাব্বুর মনে হতে থাকে তার সামনে সবকিছু ঘুরছে, তার সাথে
সাথে সে নিজেও ঘুরছে। মনে হতে থাকে সে এক্ষুনি বুঝি উল্টে পড়ে
যাবে, তাই সে শক্ত করে টেবিলটা ধরে রাখল যেন তারা পড়ে না যায়।
গাব্বু চোখ বড় বড় করে তাকিয়ে রইল, সবাইকে কাপসা কাপসা দেখাচ্ছে,
আস্তে আস্তে সবাই স্পষ্ট হতে শুরু করল।

সবাই গাব্বুর দিকে একদৃষ্টে তাকিয়ে ছিল, মাথায় উস্কখুস্ক চুল, চোখে
৪পর ফ্রেম্ চশমা, আব্বু গম্ভীর গলায় জিজ্ঞেস করলেন, "গাব্বু,
এখানে কী হচ্ছে বলবি?"

গাব্বু হাসি হাসি মুখে বলল, "একটা এক্সপেরিমেন্ট করলাম।"

"সেটা তো বুঝতেই পারছি। কিন্তু কিসের এক্সপেরিমেন্ট?"

গাব্বু সবার দিকে তাকাল, তারপর নিশ্বাস ফেলে বলল, "তোমরা
মনে হয় বুঝবে না।"

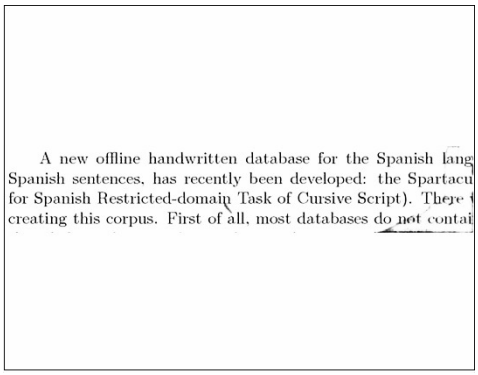
গাব্বুর বয়স মাত্র বারো, কিন্তু এর মাঝেই তার চুলের ডগা থেকে
পায়ের নখ পর্যন্ত সে পুরোপুরি বৈজ্ঞানিক। সেটি কোনো বড় সমস্যা
হওয়ার কথা না, কিন্তু সে যেহেতু প্রতি মুহূর্তেই কোনো না কোনো
বৈজ্ঞানিক পরীক্ষা করছে, এবং তার বেশির ভাগ পরীক্ষার ধকলগুলো বাসার
সবাইকে সহ্য করতে হচ্ছে, তাই বাসার সবারই "এক্সপেরিমেন্ট" শব্দটার
সাথে একধরনের অ্যালার্জির মতো হয়ে গেছে। তাই যখন গাব্বু তার
বৈজ্ঞানিক এক্সপেরিমেন্টের কথা সবাইকে বোঝানোর চেষ্টা করে অন্যেরা
অনেক সময়ই তার গুরুত্বটুকু গাব্বুর মতো করে বুঝতে পারে না।

আব্বু মুখে গাম্ভীর্য ধরে রেখে বললেন, "চেষ্টা করে দেখ, বুঝতেও তো
পারি।"

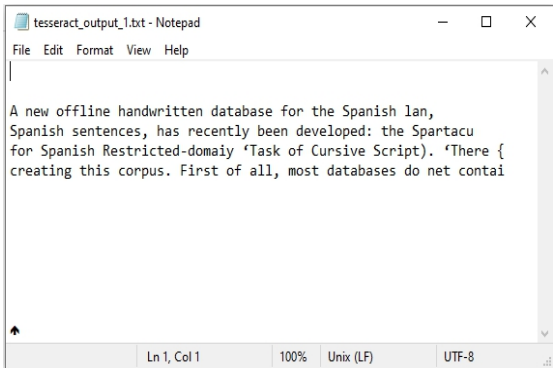
একটা বৈজ্ঞানিক বিষয় বোঝানোর সুযোগ পেয়ে এবার গাব্বুর চোখ-
মুখ ঝলমল করে ওঠে, সে চোখ বড় বড় করে বলল, "তোমরা সবাই
নিশ্চয়ই ভাবো যে, কান দিয়ে আমরা শুধু শুনি। এটা সত্যি না।"

১০

Figure 34: Clean image **Figure 35: OCR for Shulikhan** **Figure 36: OCR for Tesseract**



A new offline handwritten database for the Spanish lang
Spanish sentences, has recently been developed: the Spartacu
for Spanish Restricted-domain Task of Cursive Script). There
creating this corpus. First of all, most databases do not contai



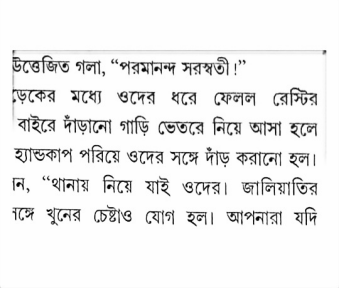
tesseract_output_1.txt - Notepad

File Edit Format View Help

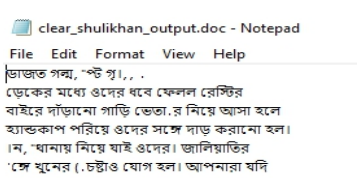
A new offline handwritten database for the Spanish lan,
Spanish sentences, has recently been developed: the Spartacu
for Spanish Restricted-domaiy 'Task of Cursive Script). 'There {
creating this corpus. First of all, most databases do net contai

Ln 1, Col 1 100% Unix (LF) UTF-8

Figure 37: low resolution clean image **Figure 38: OCR for Tesseract**



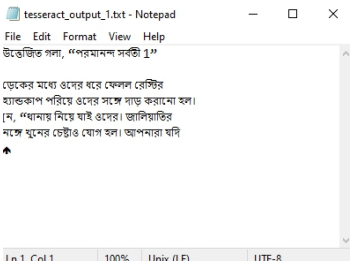
উত্তেজিত গলা, "পরমানন্দ সরস্বতী!"
ডেকের মধ্যে ওদের ধরে ফেলল রেস্টির
বাইরে দাঁড়ানো গাড়ি ভেতরে নিয়ে আসা হলে
হ্যান্ডকাপ পরিয়ে ওদের সঙ্গে দাঁড় করানো হল।
ন, "থানায় নিয়ে যাই ওদের। জালিয়াতির
ঙ্গে খুনের চেষ্টাও যোগ হল। আপনারা যদি



clear_shulikhan_output.doc - Notepad

File Edit Format View Help

ডাজত গল্া, "ট গ।, , .
ডেকের মধ্যে ওদের ধবে ফেলল রেস্টির
বাইরে দাঁড়ানো গাড়ি ভেতা.র নিয়ে আসা হলে
হ্যান্ডকাপ পরিয়ে ওদের সঙ্গে দাড় করানো হল।
।ন, "ধানায় নিয়ে যাই ওদের। জালিয়াতির
'ঙ্গে খুনের (.চষ্টাও যোগ হল। আপনারা যদি



tesseract_output_1.txt - Notepad

File Edit Format View Help

উত্তেজিত গলা, "পরমানন্দ সরস্তী 1"

ডেকের মধ্যে ওদের ধরে ফেলল রেস্টির
হ্যান্ডকাপ পরিয়ে ওদের সঙ্গে দাড় করানো হল।
।ন, "ধানায় নিয়ে যাই ওদের। জালিয়াতির
নঙ্গে খুনের চেষ্টাও যোগ হল। আপনারা যদি

Figure 39: Clean image **Figure 40: OCR for Shulikhan** **Figure 41: OCR for Tesseract**

4.3 OCR Accuracy

We got different accuracy for different situation. We calculated all the accuracy based on correct word count from OCR. We didn't consider trained size images accuracy because it is very similar to the previous systems results. So we only showed the improved accuracy where our approach shines. Thus we calculated and showed only the accuracy for low and high resolution images.

Type of OCR / System	Shulikhan OCR Accuracy		Tesseract OCR Accuracy	
Size	Low	High	Low	High
Without Pre-processing System	45.2%	40.3%	35.7%	51.1%
Our System	80.9%	98.4%	83.3%	98.7%

Table 10: OCR Accuracy

4.4 Future Work

We have come up with a solution that works in every pre-processing method. We only implemented it in CNN Autoencoder. It can also be useful for other denoising method also. With the resources we had, we did everything to improve the OCR result. But we think our method can be very effective with much more resources. Though our method works for small and large images, OCR accuracy of small sized images is not the highest. We think that resolution enhancement of small sized images can result in more accurate OCR output.

Chapter 5

CONCLUSION

Our approach improves OCR results significantly. Our approach made a huge jump in accuracy for higher and lower resolution images alongside with trained sized images. We got around 98.7% accuracy on word count in large scale images and 83.3% accuracy on small sized images. Image processing field is much larger and is improving day by day and we think our work will leave a great impression. Zero padding, cropping, reconstruction of images is not only useful for our approach but also these can improve the results of other such approaches.

References

[1] Quang Anh Bui, David Mollard, and Salvatore Tabbone, (2017). "Selecting automatically pre-processing methods to improve OCR performances," IAPR International Conference on Document Analysis and Recognition (ICDAR).

[2] Brij Mohan Singh • Mridula, (2014). "Efficient binarization technique for severely degraded document images" , CSI Publications 2014.

[3] A. El Harraj[1] and N. Raissouni[2,] (2015). "OCR ACCURACY IMPROVEMENT ON DOCUMENT IMAGES THROUGH A NOVEL PRE-PROCESSING APPROACH", Signal & Image Processing An International Journal.

[4] Ines Ben Messaoud, Hamid Amiri, Haikal El Abed, Volker Margner,(2012). "Document Preprocessing System - Automatic Selection of Binarization", IAPR International Workshop on Document Analysis Systems.

[5] Ekta Vats , Anders Hast , Prashant Singh,(2017). "Automatic Document Image Binarization using Bayesian Optimization", Proceedings of the 4th International Workshop on Historical Document Imaging and Processing.

[6] Huimin Lu[1] , Baofeng Guo[1], Juntao Liu[2], Xijun Yan[2] (2017). "A shadow Removal Method for Tesseract Text Recognition", International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI).

[7] J. Pastor-Pellicer,S. España-Boquera,F. Zamora-Martínez,M. Zeshan Afzal,Maria Jose Castro-Bleda,(2015). "Insights on the Use of Convolutional Neural Networks for Document Image Binarization", International Work-Conference on Artificial Neural Networks(IWANN) 2015

[8] Subhajit Chaudhury, Hiya Roy (2017). "Can fully convolutional networks perform well for general image restoration problems?", 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)

[9] Ankit Lat, C. V. Jawahar,(2018). "Enhancing OCR Accuracy with Super Resolution", International Conference on Pattern Recognition (ICPR)

[10] Cleaning Up Dirty Scanned Documents with Deep Learning, Available : https://medium.com/illuin/cleaning-up-dirty-scanned-documents-with-deep-learning-2e8e6de6cfa6

[11] DIBCO Datasets, Available : https://vc.ee.duth.gr/dibco2019/?fbclid=IwAR0rhqswSMLm36clEq2mDeRaASddSETLIWP4smmGlM-WL3TdgUoV-iZBNFk

[12] Kaggle dataset , Available : https://www.kaggle.com/aakashnain/denoising-autoencoders-to-the-rescue

[13] Our Datasets , Available : https://drive.google.com/file/d/1D3N6XrBr4TEnv2skj1hDJjGgHBeBUufN/view?usp=sharing