2nd National Level Students' Research Conference on
"Innovative Ideas and Invention with Sustainability in Computer Science and IT-2021"
In association with International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Digitization of Data Using OCR

Rutwik Shete
School Of Computer Science, MIT WPU, Pune, Maharashtra, India

ABSTRACT

In this modern world we hear buzzwords like Artificial Intelligence and Machine Learning whose application in the tech industries not only mesmerises us but creates an important landmark on human minds. Interestingly the second part of both the words, that is intelligence and learning respectively , are quite entangled with each other. They emphasise on the importance of the past data. As we all learn from the data our ancestors produced and we are creating new for our future generations. Unfortunately our ancestors could not keep that data on computer or on the cloud due to lack of resources. Instead, they put it on rocks and paper. This paper is an attempt to develop a system which will digitize the data on paper to be consumed in Machine Learning Models to achieve better precision in predictions. This system starts with just clicking a clear photo of a bill / printed document / invoice or any data on paper. Then it will be preprocessed for better end results by adjusting its saturation, brightness and other characteristics. This will then allow us to go further and call the Google's Vision AI API (Application Program Interface) which has the capability to read the document and return back the text which may or may not be in a linear fashion. Hence this text needs to be post processed in a way in which it could be further used for storing or utilizing it in the Machine Learning Models.

Keywords : Artificial Intelligence; Machine Learning; Data; Digitize; Preprocessed; Google Vision AI API; Machine Learning Models.

I. INTRODUCTION

Welcome to the era of Data , where you will hear the words like data collection , data analysis, data mining, data preprocessing and many more.. It is no surprise to us anymore that the use of the historical data in prediction helped many industries and businesses to predict their cash flow, sales, production, identify potential customers and their credibility; the list is ever increasing. However, data does not process itself, neither does it find patterns in itself. This is where Artificial Intelligence and Machine Learning swoop in to save the day, yet they are nothing without that data. Reason is simple: the more the data, the more precise the prediction is. For example, predicting the sales of mobile on a monthly basis may be easy if we have last year's sales report however predicting today's weather might not be precise even after consuming years of weather reports. Many business functions like Cash Flow are complex to predict and demand years of data to be fairly close to reality in predictions. We humans have been creating and collecting lots of data; in past few years, for instance we have created more than 50000 Exabytes of data

which in layman's term is 7,853,654,484,114,286 (more than 7 quadrillion) images in the year 2020 itself. Now this data consists of photos / bills & invoices / search engine data / production data so on and so forth. There are some business cycles which can't be predicted in a few years' data which demands digitization of the data that still exists on paper and has no digital address of its own . The Indian government has taken a lot of efforts to digitize our revenue department where all the lands are digitally mapped to something called 7/12. We can imagine the efforts of reading each of the transcripts and digitizing it with human efforts. OCR has provided us with a choice now either to get a team of trained personals to work day and night to read the transcripts and digitize them or use AI and Machine Learning so it will literally be like the Idiom in Hindi "Lohe Ko Loha Katta Hai".

II. PROPOSED SYSTEM

We are at a point where OCRs (Optical Character Recognition) have already been constructed by the current tech giants which provide us the text from the images. With the help of these OCR APIs we can fetch the plain text quite precisely. However, results are a function of quality of the image. To achieve better results we shall use some basic but important preprocessing techniques such as (A) Binarization, (B) Skew Correction, (C) Noise Removal, (D) Thinning & Skeletonization. After correcting the image in all possible ways we will be sending it to the Google Vision AI API which is a cloud base OCR provided by google which further provides us with the set of texts that is detected by the OCR. As these texts are not perfectly detected in the order as they appear in the photo , we will have to go through post processing to achieve the similar arrangement of the sentences and words[2]. "Et Voila ! ", you have digitized the photo . The following system can be further improved and advanced to make it capable of reading a human written handwriting on the bills and prescription.

III. FLOW CHART & FIGURES

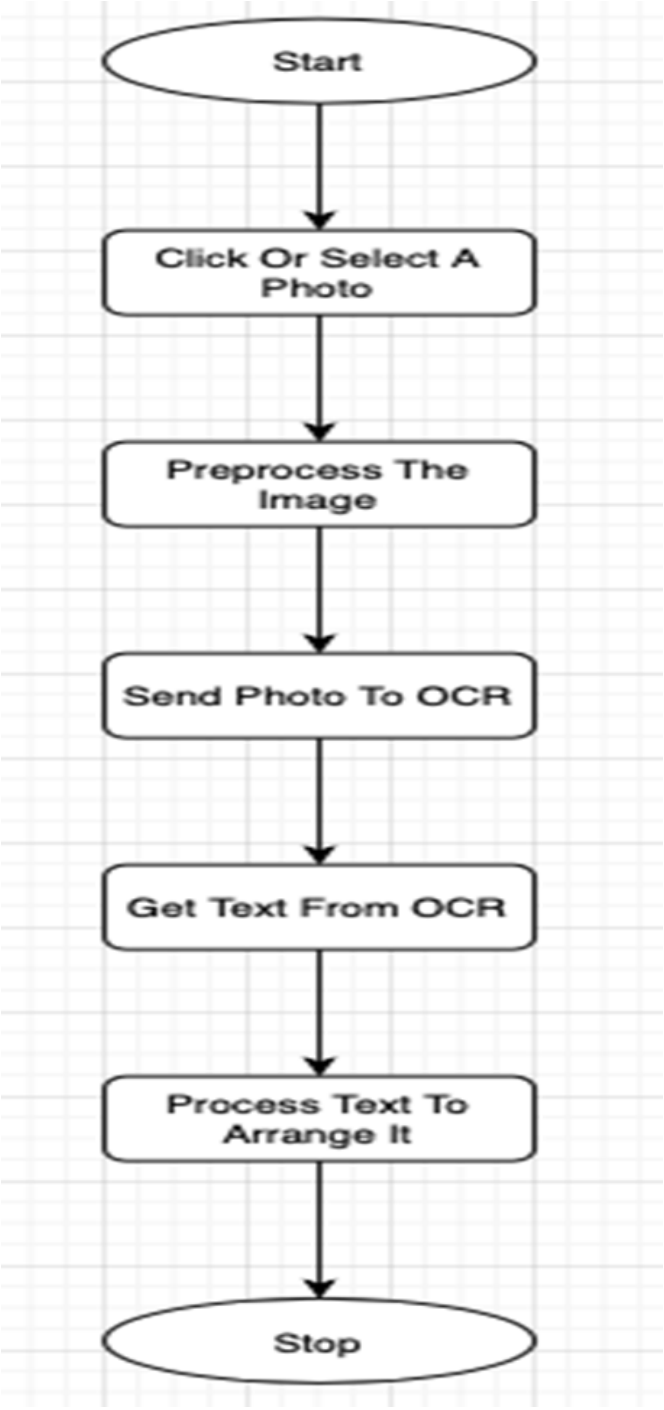

Figure 1.0

Figure 1.1

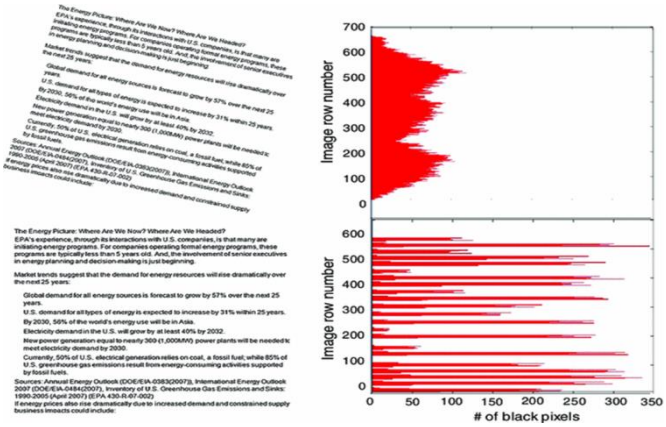

Figure 1.2

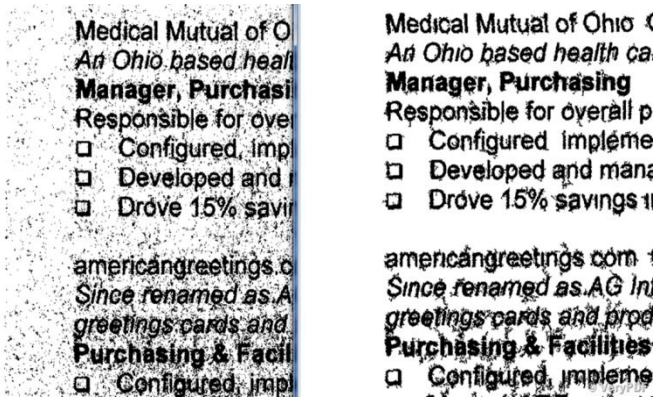

Figure 1.3

Figure 1.4

IV. WORKING

A. Getting Started (Figure 1.0)

The system of digitizing starts with a simple step of clicking / selecting / scanning an image of whichever paper document you wanna digitize. Then comes the preprocessing whose techniques we will be looking into after a few lines. Now that taken care of we send the image to any Optical Character Recognition (OCR) system which may be cloud based or in house developed system. This will then send us text back in return which may or may not be arranged in a desirable arrangement. So to achieve that we will have to do some processing on the text to get the favourable output. This is how we will digitize your data from a photo of a paper in no time and without putting as much cost and time as you would put in for a person to sit and enter the data manually.

B. Image Preprocessing

As mentioned above regarding preprocessing there are some most basic and important techniques we will be going through.

- Binarization (Figure 1.1) : This Method in layman's term is to make the image into a black and white image. Black pixel value = 0 and white pixel = 255 . The threshold is considered as 127 as it is the midpoint of range 0 - 255 . If the pixel is greater than threshold it will be considered as white pixel, else considered black pixel[3].
- Skew Correction (Figure 1.2) : While clicking a photo it may be slightly skewed (Having an oblique or slanting direction or position). To overcome this skewness is very important to get the text in a proper sequence. There are many methods like (a) Projection Profile Method , (b) Hough Transformation Method , (c) Topline Method, (d) Scanline Method[4]. For our application we will be going with the first

method Projection Profile Method which is the easiest and most widely used.

- Noise Removal (Figure 1.3) : Noise removal is used to soften the image by removing the dots / patches which have a higher contrast than the rest of the image. This process can be performed on both the colored and binary images[5].
- Thinning & Skeletonization (Figure 1.4) : This one method is an optional process . If we are using the OCR to read printed text then there is no need of thinning but if we are using it to detect a handwritten text then, due to diverse styles and different stroke width while writing we will have to implement thinning and skeletonization on it[6].

V. COSTING

The cost of digitizing 10 photos is as less as buying a ₹1 chocolate. For every month you get 1000 images free and then the rate of digitizing 1 photo defers between $0.0006 - $0.0015[7].

VI. CONCLUSION

Data insufficiency is a major game changer for precise predictions by any Machine Learning Models. Our current business functions being global became ever complex and needed assistance from AI/ML to perform predictions for them to make meaningful decisions. With advancement of OCRs technologies we can help businesses by providing them with their historical data available for predictions. And it can be achieved in seconds and at very nominal cost. Further we can train the OCR to read even hand written scripts with the same accuracy and speed, So sky is the only limit for OCR and how it could help confidence level of predictive models!

VII. REFERENCES

[1]. The Cambrian Data Explosion

[2]. Google Vision Ai OCR API

[3]. Jyotsna, S. Chauhan, E. Sharma and A. Doegar, "Binarization techniques for degraded document images — A review," 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, 2016, pp. 163–166, doi: 10.1109/ICRITO.2016.7784945.

[4]. A. Papandreou and B. Gatos, "A Novel Skew Detection Technique Based on Vertical Projections," 2011 International Conference on Document Analysis and Recognition, Beijing, 2011, pp. 384–388, doi: 10.1109/ICDAR.2011.85..

[5]. K. Lin, T. H. Li, S. Liu and G. Li, "Real Photographs Denoising With Noise Domain Adaptation and Attentive Generative Adversarial Network," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 2019, pp. 1717–1721, doi: 10.1109/CVPRW.2019.00221.

[6]. Choudhary, Amit & Rishi, Rahul & Savita, Ahlawat. (2013). A New Character Segmentation Approach for Off-Line Cursive Handwritten Words.Procedia Computer Science. 17. 88–95. 10.1016/j.procs.2013.05.013.

[7]. Google Vision Api Pricings