F.Y.B.Tech Students' Conference on Capstone Projects 2 on 7th and 8th August 2021, Vishwakarma Institute of Technology, Pune, INDIA.

Image to Text Conversion Using Machine Learning

Kiran More, Abu Dujana, Adarsh Acharya, Priyanka Acharya, Shruti Adhau, Swapnil Adhav

Department of Engineering, Sciences and Humanities (DESH)

Abstract *— Today the most information is available either on paper or in the form of scanned documents, Portable Document Formats (PDF). To perform operations, this information must be either in text format or editable format. Extracting data from scanned documents is not an easy task. Writing down the information from an image can be very tedious, if the information is large. Thus, the system is needed to extract text from general backgrounds. So, the projects aim to create an Optical Character Reader (OCR) which extract data from images and converts it into editable format. The proposed system is successful in recognizing handwritten digits from MNIST DATABASE. Extracted data can be further converted into a Word Document or simple Text File.*

I. INTRODUCTION

Image processing, within the computer vision is could be improving image quality, resolution, extracting important information from the image. With the advancement in the technology, there are many new techniques for manipulating photographs, images and Pdfs. The text recognition is of much importance in many areas. But it's a tedious task for a machine. For recognition, we've to train the system to identify the text. The character recognition involves several steps like image acquisition, pre-processing, segmentation, feature extraction, training a Neural Network and post-processing. The prime aim of the project is to develop a system that may efficiently recognize text from a Pdf and convert it into the format of user's choice. The system is currently restricted in recognizing characters from clean and clear backgrounds. Further the accuracy decreases, as system confuses with similar characters with dim backgrounds. But the system is perfect for clear images.

II. LITERATURE REVIEW

[1] Chandrahas Gaikwad, Satish Akolkar, Reshma Khodade- 'Generic PDF to Text Conversion using Machine Learning' (2014): Presented a generic way of making PDF documents editable by the script-independent and machine learning features. Implemented decision model that systemises the classification of characters identified from the pdf.

[2] Noman Islam, Zeeshan Islam, Nazia Noor- 'A Survey on Optical Character Recognition System' (2016): Summarize the research so far done in the field of OCR. Provided an overview of different aspects of OCR and discusses corresponding proposals aimed at resolving issues of OCR. Discussed acquisition, pre- processing, segmentation, feature extraction, classification and post-processing in detail.

[3] Polaiah Bojja, Naga Sai Satya Teja Velpuri- 'Handwritten Text Recognition using Machine Learning Techniques in Application of NLP' (2019): Built a Model to analyse the text written and convert it in Computer Text and Voice formats. Identify input characters or image correctly then analysed to many automated process systems. Successfully, recognised text from input image with an accuracy of 92.7%.

[4] Nisha Pawar, Zainab Shaikh, Poonam Shinde, Prof. Y.P. Warke- 'Image to Text Conversion Using Tesseract' (2019): Proposed the system that can be used for character recognition from scanned documents so that data can be digitalized. Also, the data can be converted to audio form so as to help visually impaired people obtain the data.

[5] Sri. Yugandhar Manchala, Jayaram Kinthali, Kowshik Kotha- 'Handwritten Text Recognition using Deep Learning with TensorFlow' (2020):

Process the input image, extraction of features, and classification schema takes place, training of system to acknowledge the text. Recognized text with accuracy of 90.3% and above.

[6] AMSHED MEMON, MAIRA SAMI, RIZWAN AHMED KHAN, MUEEN UDDIN- ' Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic': Summarize research that has been conducted on character recognition of handwritten documents and provide research directions. Synthesized and analysed research articles on the topic of handwritten OCR (and closely related topics) which were published between year 2000 to 2019. Extracted and analysed research publications on six widely spoken languages.

[7] Tien Fui Yong, Saiful Azad, Mohammed Mostafizur Rahman, Kamal Z. Zamli - 'A Highly Accurate PDF-to-Text Conversion System for Academic Papers using Natural Language Processing Approach': Proposed a Natural Language Processing based PDF-to-text (NLPDF) conversion system, which comprises of reads contents from the PDF and reconstruct the text. It can extract word-based tokens well.

[8] Sahana K Adyanthaya – 'Text Recognition from Images: A Study (2020)': The paper provides a good summary of all the steps used in text recognition and extraction from images. The previously available work carried out in the field of text recognition has also been discussed briefly. A review of basic flow of text recognition from images has been mentioned. Finally, the uses of text recognition in various fields are discussed.

[9] K. Karthick, K.B. Ravindra Kumar, R. Francis, S. Ilankannan- 'Steps Involved in Text Recognition and Recent Research in OCR; A Study': The study found that the segmentation free approach is possible in OCR using Deep Neural Networks (DNN). It was noticed that multilingual character segmentation and recognition is possible with better rate and less computation time. This helps in obtaining optimal results.

[10] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, 'EMIST: an extension of mnist to handwritten letters': This paper introduced the EMNIST datasets which is a suite of six datasets. The intension of EMIST to provide a more challenging alternative to the MNIST dataset. The characters of the NIST Special Database 19 were converted to a format that matches that of the MNIST dataset. That made it immediately compatible with any network capable of working with the original MNIST dataset.

***Tools*: -**

a) Python interpreter.

b) Visual Studio Code.

c) Machine Learning Algorithm.

d) Python libraries.

III. WORKING

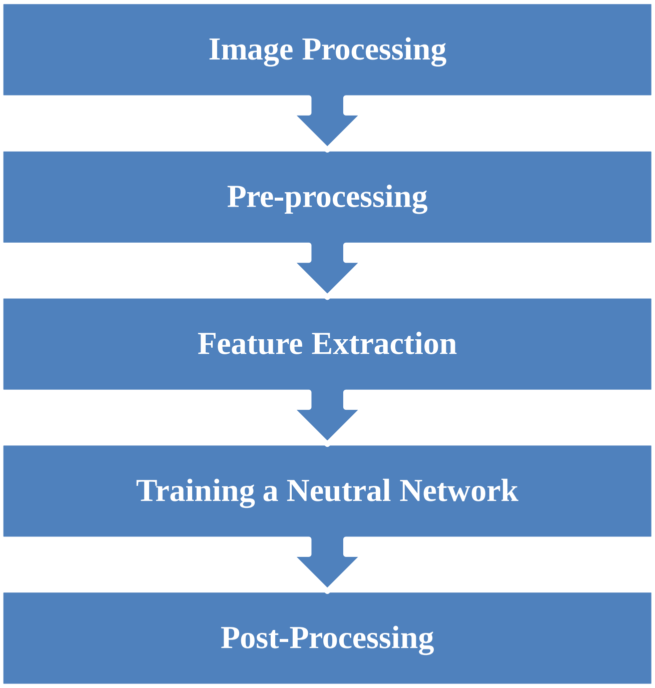

Fig.1 Block diagram

Every OCR works in 6 basic steps

3.1. Image Acquisition

First, an image is scanned or captured and stored, then using image processing and image segmentation, it converts that image into black and white by replacing each pixel with particular shade of black and white.

3.2. preprocessing

To make the data readable for computers, the noise level of the image is optimized and the extra area around the texts are removed. This is especially vital when dealing with handwritten texts.

3.3. Segmentation:

The image is then sliced line by line and then character by character to make them meaningful for the computer.

3.4. Feature extraction

This step means splitting the input data into a set of features, which means to find essential characteristics that make one or another pattern recognizable. As a result, each character gets classified in a particular class.

3.5. Training a Neural Network

Once all features are extracted, they are passed into a set of neural networks to train it to recognize different characters. The better the training data set is, the better the OCR works.

3.6. Post-processing

This part comes after the recognition of texts. if possible, few minor spelling errors get corrected.

IV. Limitations

OCR methods require high-quality or high-resolution pictures with great precision for good character recognition. Several fundamental structural features, such as significant text differentiation as well as the background. It is crucial to consider how images are created, since it is a deciding factor in the accuracy and success of OCR. This has a significant impact on image quality. OCR using pictures produced by scanners usually yields excellent accuracy and precision. Excellent performance Camera pictures, on the other hand, are a different story are not always as good as scanned photos as a source of information. Due to environmental or camera-related variables, OCR is not possible. There might be a slew of mistakes, all of which are explained here.

4.1 Scene Complexity

In a typical setting, we may observe a significant number of man-made items, such as paintings, buildings, and symbols, that are included in camera-captured photos. Because these items have similar structures and appearances to text, text detection in the processed image is difficult. The text is formatted in a way that makes it easy to read.

4.2 Conditions of Uneven Lighting

When photographing in natural settings, uneven lighting and shadows are common. This is a problem for OCR since it weakens the image's desirable properties, resulting in less accurate detection, segmentation, and recognition outcomes. This characteristic of uneven illumination is what separates a scanned image from one captured with a camera. Scanned images are favored over camera photos because they have superior features and quality due to the lack of such discrepancies in lighting and shadows. Although utilizing an on-camera flash might help solve difficulties like uneven illumination, it also comes with its own set of drawbacks.

4.3 Skewness (Rotation)

In optical character recognition systems, the point of view for the input picture, which is acquired from a camera on a hand-held device or other devices, is not fixed like it is in a scanner, therefore skewing of text lines from their unique alignment may occur. When such a skewed image is submitted to the OCR classifier, the results will be extremely bad. Many approaches exist for the goal of deskewing picture documents, including Projection Profile, RAST algorithm, Hough transform, Fourier transformation methods, and so on.

4.4 Blurring and Degradation

Since operating over a selection of distances are supposed to varied digital cameras, associate vital issue is the digital cameras focusing. For the simplest accuracy of character recognition and character segmentation, character sharpness is required. At massive apertures and short distances, uneven focus is discovered once a little purpose of read changes. For the foremost half connected with photography, there are 2 styles of obscure that is: out of focus obscure and movement obscure. At the purpose for catching a moving item, once the shade rate of the camera is not sufficiently high, the device gets conferred to a regularly dynamic scene. Accordingly, blurring can discover in components in motion.

4.5. Aspect Ratios

Text has completely different side ratios. Text could also be temporary equivalent to traffic signs, whereas alternative text could be a lot of longer, such as video captions. Location, scale and length of text got to be thought about with search procedure to discover text, that introduces high machine complexity.

4.6 Tilting (Perspective Distortion)

Document pictures obtained by scanners is consistently parallel to the plane of sensor, however this cannot be discovered all times for recorded image obtained by a moveable camera, that might not typically be parallel to the shape plane. Accordingly, lines of text that distant from the camera appear littler than people who nearer to the camera that appears greater. This scenario causes tipped pictures. Observation of a perspective distortion is clear if the acknowledger is not perspective intolerant, that causes lower recognition rate and accuracy. Cell phones have a bonus with orientation sensors. They can recognize whether the device is tipped and once twisting happens they will forbid shoppers to require pictures. allowing the user to align plane of the shape with that of the camera is conjointly provided by this feature. Orientation sensors so might assure that made footage satisfy a precise degree of evenness.

4.7 Fonts

Italic vogue and script fonts of characters would possibly overlap every other character, making it troublesome to perform some of the main OCR processes. Characters of assorted fonts have giant within-class variations and kind several pattern sub-spaces, making it difficult to perform correct recognition once the character category variety is large.

4.8 Bilingual Environments

An enormous portion of the languages of Latin have several characters, languages for example, Japanese, Chinese and Korean, have a large range of character classes. Connected characters are existing in Arabic languages, that according to context, it changes writing shape. In Hindi, syllables are represented by combining letters into thousands of shapes. In bilingual situations, OCR in scanned documents stays as a primary analysis issue, since OCR in complicated symbolism is additionally troublesome.

4.9 Warping

Content or text on objects of varied geometries is another. challenge for OCR to be recognized once pictures of such. scenario captured by hand-held cameras. A few circumstances might emerge with flatbed scanners, whereby the twisted text. determined when the content procured on picture, as an example the content towards the binding of a very thick book. For convention paper documents, a technique for image dewarping is planned by Ulges Et Al. By expecting the method that content lines are equally separated and parallel to every other, they dewarp pictures.

V. Future Scope

Data conversion is an important part of any business, which requires converting one form of data into another useable format. To overcome this problem, we can successfully use this and convert the data from one form to another easily. For e.g.: IMG to TEXT And it is more useful to convert PDF to TEXT, more over Today the most information is available either on paper or in the form of scanned documents, Portable Document Formats (PDF). The conventional PDF to text conversion software is incapable of editing some unexplored scripts. In this project, a generic way of making PDF documents editable by the script-independent and machine learning features is presented. This is possible by cropping out the characters from the PDF. A set of classifiers is applied to identify characters. The Decision Model implemented as a part of Machine learning organizes the classifier functions. The resultant classifier set gives the resolution for the character. This approach removes the barrier of limiting our scope to international scripts and also facilitates usage of regional scripts in the technological world. This is how we can also convert regional scripts.

VI. Conclusion

1. This project results in saving time of the user.

2. Tedious way of seeing and writing the information gets solved within just few clicks using this project.
3. All sort of typed information will be accurately converted to text format.

VII. Acknowledgment

It gives immense pleasure to express our deep sense of gratitude with sincere thanks and appreciation to our project guide Prof. Kiran More Sir for suggesting and supporting us to carry the project work. We would also like to thank Prof. Mahajan Sir, for his valuable guidance. We were able to complete this project successfully with the help of valuable guidance and constant encouragement of them.

References

[1] Chandrahas Gaikwad, Satish Akolkar, Reshma Khodade, Deepali, Dalal Swarupa Kamble - Generic PDF to Text Conversion using Machine Learning, International Journal of Computer Applications (0975 – 8887) Volume 106 – No. 12, November 2014.

[2] Noman Islam, Zeeshan Islam, Nazia Noor - A Survey on Optical Character Recognition System, Journal of Information & Communication Technology -JICT Vol. 10 Issue. 2, December 2016.

[3] Polaiah Bojja, Naga Sai Satya Teja Velpuri, Gautham Kumar Pandala, S D Lalitha Rao Sharma Polavarapu, Pamula Raja Kumari - Handwritten Text Recognition using Machine Learning Techniques in Application of NLP' (2019), International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume -9 Issue -2, December 2019.

[4] Nisha Pawar, Zainab Shaikh, Poonam Shinde, Prof. Y.P. Warke- Image to Text Conversion Using Tesseract, International Research Journal of Engineering and Technology, (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 02| Feb 2019 www.irjet.net p-ISSN: 2395-007.

[5] Sri. Yugandhar Manchala, Jayaram Kinthali, Kowshik Kotha, Kanithi Santosh Kumar, Jagilinki Jayalaxmi - Handwritten Text Recognition using Deep Learning with TensorFlow, International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV9IS050534 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by: www.ijert.org Vol.9 Issue 05, May 2020.

[6] JAMSHED MEMON, MAIRA SAMI, RIZWAN AHMED KHAN, MUEEN UDDIN - Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic, Received June 24, 2020, accepted July 16, 2020, date of publication July 28, 2020, date of current version August 14, 2020. Digital Object Identifier 10.1109/ACCESS.2020.3012542.

[7] Tien Fui Yong, Saiful Azad, Mohammed Mostafizur Rahman, Kamal Z. Zamli - A Highly Accurate PDF-to-Text Conversion System for Academic Papers using Natural Language Processing Approach, Copyright © XXXX American Scientific Publishers Advanced Science Letters All rights reserved Vol. XXXXXXXXX Printed in the United States of America.

[8] Sahana K Adyanthaya-Text Recognition from Images: A Study, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org NCCDS - 2020 Conference Proceedings.

[9] K. Karthick, K.B. Ravindra Kumar, R. Francis, S. Ilankannan, Steps Involved in Text Recognition and Recent Research in OCR; A Study, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019.

[10] [10] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "Emnist: an extension of mnist to handwritten letters," arXiv preprint arXiv:1702.05373, 2017.