OPTICAL CHARACTER RECOGNITION IN HEALTHCARE

Dr. Deepali Ujalambkar[2], Adarsh Bhosale[1], Dipti Shegar[1], Aditya Tandulwadkar[1], Purva Wagh[1]

[2]Dr. Deepali Ujalambkar, [1]Adarsh Bhosale, [1]Dipti Shegar, [1]Aditya Tandulwadkar, [1]Purva Wagh
[1]Department of Computer Engineering, AISSMS College of Engineering, Pune, India
[2]Assistant Professor, Department of Computer Engineering, AISSMS College of Engineering, Pune, India

---

Abstract - *In this paper, we have proposed an Optical Character Recognition (OCR) model which is also called as Text Recognition Technology using Machine Learning and Deep Learning for effectively extracting text information from medical records, claims, EOBs, or any other medical forms. This will speed up the accessibility of medical records or healthcare information and will also wouldn't have any errors. This will ensure that data accessed is available easily to analyse and will be available for wide assortments of medical records and forms within 24 hours. The challenge here is to develop an application which will recognize characters in the medical form/record images fed to it on computer system when information will be scanned through images it'll convert all paper medical records to electronic format. Input could be given to this model either through a digital image or scanning printed text. This model could be also extended to acknowledge alphabets from various languages. Through optical character recognition you can effortlessly digitize a document without any need for manual data entry. This's the reason OCR is commonly utilize for business flow optimization and automation. In this paper we also present the survey of different methodologies and algorithms used in previously published papers.*

Key Words: Machine Learning, Deep Learning, Optical Character Recognition, Character Recognition, Handwritten Character Recognition.

1. INTRODUCTION

Optical Character Recognition of machine characters is a growing area of research and it has got numerous applications in various sectors like companies, offices, industries and banks etc. This paper represents a simple and also portable approach for optical character recognition based system which would recognize the characters from given image using Machine Learning and Deep Learning. Many medical claims are processed every year, which could create plenty of paperwork and also manual processing. This work is then further heighten if records are missing or erroneous. There has been a growing industry-wide push since past several years to convert all paper medical records to electronic format. OCR ushers fewer manual processes, which then allows more efficiency and accuracy. The objective of this paper is to present an expert system for Optical Character Recognition using Neural Network, which could effectively apprehend a particular character using

Artificial Neural Network(ANN) approach. The output text obtained from this OCR is further used for digital document editing, and compact data storage and also available for future analysis and insights.

2 RELATED WORK

- Anchal Garg and Rishabh Mittal [1] in this paper introduces the idea, explains the method of extraction, and presents the most recent techniques, technologies, and current analysis within this area. Such a review can facilitate different researchers within this field to get a summary of the technology. They performed text extraction from historical Documents and from television. Their system needed filtered quality images and handwritten images record such as near oblique, touching/overlapping content lines.

- Prof. Dr. Pravin R. Hutane, Mrunal G. Marne [2] in this paper describes Tesseract, Tesseract is one of the most widely use open source libraries for implementing OCR in Android applications. A tesseract is free software that is originally developed by Hewlett and Packard. This tesseract stands out from all accessible OCR engines. Also tesseract is an open-source library that can be easily integrated into Android applications. Memory management is a difficulty with tesseract and it does not have a large database.

- Dr Sunanda Dixit1, Bharath M2, Amith Y2, Goutham M L2, Ayappa K2, Harshitha D2 [3] in this paper, the accessibility of the Optical Character Recognition system to the visually impaired is a great progress, Where users can scan books, incoming mail or other programs. Aim of this paper at presenting an OCR utility that recognizes text characters with the help of a machine-learning model. This model is able to recognize the text in optical forms. Using this system we can easily recognize texts in optical form. Also prediction accuracy can be increased by the training with the help of more trained images. This process is time consuming and expensive. Though the quality is not always high. Omer Aydin [4] in this document, two different models to extract the text from the images. One is OCR using Machine Learning, and the other is MODI (Microsoft Office Document

Imaging Library). They got 351 character and 62 words using MODI. Their results were 88.50% and 85.20% word match rates using MODI match and OCR code. This character matching model was found to match 340 characters using both methods for a match rate of 97.7%. The accuracy of the results was around 45% to 60%. This result is a little low. These results can be improved by using different phases. For example, you can add a dictionary that matches exact words and use methods for that.

- Ilanchezhian P1, Jayanthi D2, Kavipriya P3, Kavin Sagar S4 [5] in this paper, they give some information about Artificial intelligence (AI) and Optical Character Recognition (OCR). They use Optical Character Recognition (OCR) to detect the text, involve in certain matching algorithms, and display the results as descriptive text and augmented reality (AR). Augmented reality (AR) provides a 'virtual visual treat' on the basis of the real world. Compared with other classifier methodologies four systems of their project procedures display higher accuracy and the decision is made based on the weight of data in the datasets.

- Mathumitha IM1, Vinodhini B2, Jayaprakash S3, MaliniDevi R4, Sangeetha K [6] in this project, a device that is developed by the web converts an image text to speech. This device helps visually impaired people and also helps in converting hardcopy hospital data into softcopy. This device is implemented with the help of a Web Camera and System. In this document, only Devanagari text-to-speech conversion is completed for Marathi on paper text. Optical Character Recognition (OCR) and text-to-speech (TTS) are two techniques applied to reach the required output which is complex.

- Shalini Sonth, Jagadish S. Kallimani [7] in this paper proposes a novel implementation of an Optical Character Recognition (OCR) based smart book reader for the visually challenged. There is a need for a portable text reader that is affordable, portable and readily available to the community. This paper addresses the integration of a complete Text read-out system designed for the visually challenged. The OCR used in this project is Google Tesseract and the TTS employed is Pico. This system recognizes captcha, detects handwritten texts and invoice imaging. Works only for noisy images. However, it adds noise when a clear image is passed. If the words within an image are too cluttered, the Segmentation module fails to carry out segmentation with perfect accuracy.

- Kartikeya Jain, Tanupriya Choudhary and Nitbhay Kashyap in [8] this paper proposed system is implementing the OCR technology to park the vehicles in a smart way and keep track of the vehicles which are entering and leaving. The system will capture the image of the number plate of the vehicle using the OCR process and will instantly update the database. The proposed system in this paper can be used in the detection of the number plate which is of any color which is a drawback of the many systems proposed. The above proposed system can be used for recognition of the number plate of all the vehicles not only for cars which was a drawback for the system in the older system proposed.

- Jagruti Chandarana, Mayank Kapadia in [9] this paper presents an overview of feature extraction methods for character recognition. Feature extraction method selection is the only most important factor in achieving high recognition performance in character recognition systems. For different representations of the characters' different feature extraction methods are designed. When a few promising feature extraction methods have been identified, they need to be evaluated experimentally to find the best method for the given application. There are various methods of the character recognition which can be divided into the following groups as Pattern systems, Structural systems, Feature systems and Neural network systems.

- Vishwanath Bijalwan1, Vinay Kumar2, Pinki Kumari3 and Jordan Pascual in [10] this paper, we first categorize the documents using KNN based machine learning approach and then return the most relevant documents. In recent years, text categorization has become an important research topic in machine learning and information retrieval and e-mail spam filtering. It also has become an important research topic in text mining, which analyses and extracts useful information from texts. More Learning techniques have been in research for dealing with text categorization. The data set used for this paper is in the form of sgml files. We have used Reuters-21578. They were labeled manually by Reuters personnel. Labels belong to 5 different category classes, such as 'people', 'places', 'Exchange', 'Organization' and 'topics'. The total number of categories is 672, but many of them occur only very rarely. The dataset is divided in 22 files of 1000 documents delimited by SGML tags.

3. PROPOSED SYSTEM

OCR is being employed for recognizing the optical characters. The "Optical Character Recognition System" is enforced employing a Convolutional Neural Network (CNN), CNN is extremely satisfactory at learning on style within the input image, like handwriting, written text, etc. CNN may be a category of Deep Neural Networks which will acknowledge and classify explicit options from pictures and square measure wide used for analyzing visual pictures. Their applications vary from image and video recognition, image classification, medical image analysis, pc vision and language process.

4. IMPORTANT MODULES AND ALGORITHMS USED

Following may be a description of all the implementation steps, that were applied so as to realize the ultimate target of our project.

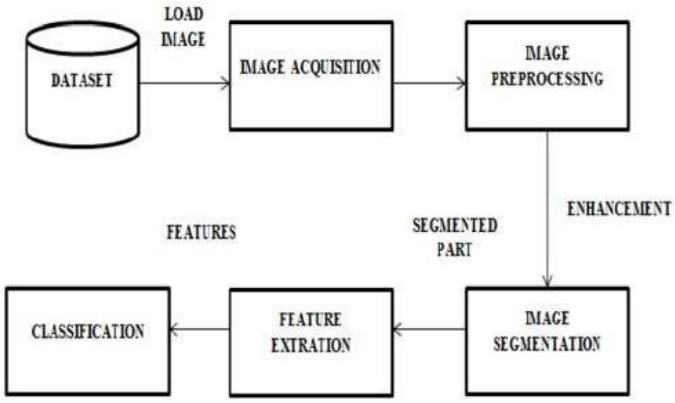

Fig.1. A Dataflow diagram

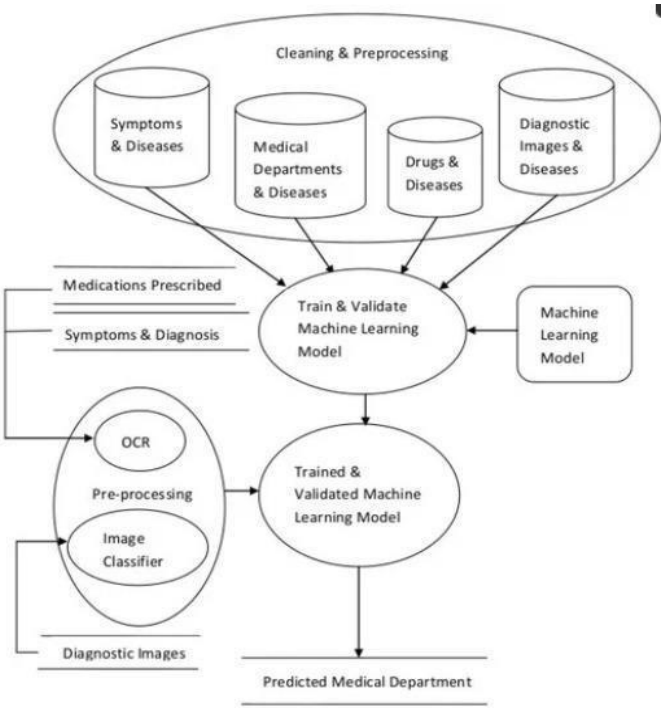

Fig.1.B Basic Block Diagram

4.1.1 Module 1: Image processing

In image preprocessing the steps enclosed are measure needed to form the input image into an appropriate type for segmentation

4.1.1.1 Grey Scale Conversion

In grey scale conversion the color image is taken as input and is regenerate into grey scale. As every color pixel is delineated by a triplet(R, G, B) for red, green and blue of intensities. we'll map all the various intensities into one variety which will provides a grey scale value.

ALGORITHIM: grey Scale Conversion INPUT: Scanned Optical Document Image

OUTPUT: grey Scaled Document Image

4.1.1.2 Image Binarization

In Image Binarization, we'll convert a picture of upto 256 grey level to a black and white image. the straightforward need to use image binarization is to decide on a threshold a price then classify them into all pixels with values higher than the edge as black and every one alternative pixels square measure white.

ALGORITHM: Image Binarization(Thresholding) INPUT: Grey Scaled Image

OUTPUT: Black and White Image

4.1.2 Module 2 : Segmentation

Once the preprocessing of image is completed then it's necessary to section the document image into lines, then lines into words, then words into the characters. when the characters has been extracted from document image, we are able to extract options from it for recognition of characters. Characters separation from the input image involves 3 steps as:

- Line Segmentation
- Word Segmentation
- Character Segmentation

4.1.2.1 Line Segmentation

While playacting line segmentation, we want to scan every horizontal pixel of rows ranging from the highest of the document. within the lines separated, wherever we discover a row with no black pixels.

ALGORITHM: Line Segmentation INPUT: Binarized Document Image

OUTPUT: segmental lines from Document Image

4.1.2.2 Word Segmentation

To perform word segmentation, we want to scan every vertical pel column ranging from the left of the road. The words square measure separated wherever we discover a column with no black pixels for quite predefined columns. This column acts as a separation between 2 words.

ALGORITHM: Word Segmentation

INPUT: segmental lines from Image and average pel breadth for word separation

OUTPUT: segmental words from line

4.1.2.3 Character Segmentation

To perform character segmentation, we want to scan every vertical pel column ranging from the left of word. The characters square measure separated wherever we have a tendency to finds a column with no black pixels columns. This column acts as a separation between 2 character.

ALGORITHM: Character Segmentation INPUT: segmental Words from lines OUTPUT: segmental Characters from words

4.1.3 Module 3 : Feature Extraction

As individual characters are separated, character pictures will be resized to fifteen x twenty pixels. If the options square measure extracted accurately then the accuracy of recognition is additional. Here we've used the fifteen x twenty means that three hundred pixels because it is for the feature vector. This extracted feature square measure keep in the .dat file.

4.1.4 Module 4: Training And Recognition

The Features extracted from previous modules are given as an input for Neural Network. The Kohonen algorithm is an automatic classification method which is the origin of Self-Organizing Maps. This SOM is used for training and recognition.

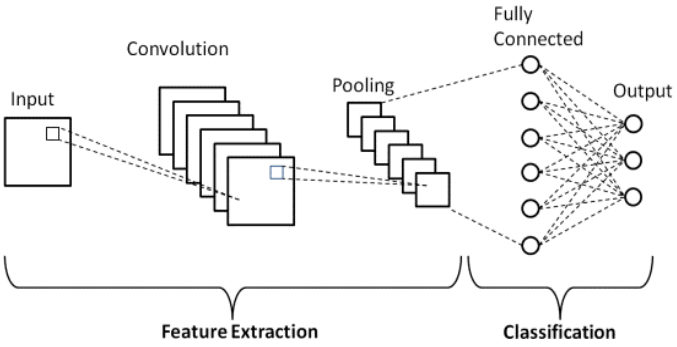

Fig.2 Convolutional Neural Network (CNN)

5. CONCLUSIONS

As we learn to recognize Optical Characters, it has been shown that recognition becomes difficult when there are odd characters or similar shapes. The scanned image is first pre-processed so the characters can be isolated from each other. Pre-processing work is performed in which normalization, filtration is executed the usage of processing steps which produce noise free and smooth output. handling our evolution set of rules with right education, assessment other step sensible procedure will cause success output of system with higher efficiency. Use of some statistical functions andgeometric capabilities through neural network will furnished higher recognition result of English characters.

REFERENCES

[1] “Text extraction with Optical Character Recognition: a systematical Review” Proceedings of the Second International Conference on the creative analysis in Computing Application (ICIRCA-2020) IEEE Xplore half Number: CFP20N67-ART; ISBN Number: 978-1-7281-5374-2.

[2] "Identifying the Optical Character Recognition Engine for Planned System" 2018 4th International Conference on Computing Communications Control and Automation (ICCUBEA).

[3] “Optical Recognition of Digital Characters Use Machine Learning” Publisher The International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), Volume 5, Issue 1, 2018, PP 9-16, Application (ICIRCA-2020) IEEE Xplore half Number: CFP20N67-ART; ISBN Number: 978-1-7281-5374-2.

[4] “Classification of Documents Extracted from Images using Optical Character Recognition Methods”, Anatolian Journal of Computer Sciences | 2021.

[5] “Patient’s Medical Report analysis using OCR and Deep Learning.”, IJCSNS International Journal of Computer Science and Network Security, VOL.22 No.3, March 2022.

[6] “OCR for Medical Data Processing”, Special Issue of Second International Conference on Advancements in Research and Development (ICARD 2021)

[7] "OCR Based Facilitator for the Visually Challenged", 2017 International Conference on Electrical, Electronics, Communication, Computer and optimization Techniques (ICEECCOT)

[8] “SMART VEHICLE IDENTIFICATION SYSTEM USING OCR", 3rd IEEE International Conference on "Computational Intelligence and communication Technology" (IEEE-CICT 2017).

[9] “A Review of Optical Character Recognition”, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 12, Dec- 2013

[10] “KNN based Machine Learning Approach for Text and Document Mining”, Vishwanath Bijalwan1, Vinay kumar2, Pinki Kumari3 and Jordan Pascual4, International Journal of Database Theory and Application Vol.7, No.1 (2014), pp.61-70