HANDWRITTEN CHARACTER RECOGNITION & TRANSLATION

Prof. Reena Deshmukh[1] Sahil Patil[2] Mahendra Rathi[3] Rahul Pawar[4]

[1, 2, 3, 4]*Department of Computer Engineering, Shivajirao S. Jondhale College of Engineering, Dombivli(E).*

[1]*reenachaudhari58@gmail.com*

[2]*sahilp2020@gmail.com*

[3]mahendra30.rathi@gmail.com

[4] *rahulpawar19012000@gmail.com*

Abstract: *HCR Using Neural Network aimed at clarifying the role of handwriting recognition in accordance with today's maturing technologies. It attempts to list and specify the elements that make up handwriting recognition and related technologies such as OCR (Optical Character Recognition) and Signature Verification [2]. Our main of the project is to make such a environment where there should be error free recognition of digit, character, the whole sentences & paragraph from handwritten document [1]. OCR has always been an intensive research topic for more than 2-3 decades. This paper will give more information about how to translate the given input data in Marathi language. It also reports about MNIST dataset for handwritten digit recognition. This database has been widely used to validate novel approaches to computer vision, and in recent years, many authors have evaluated the effectiveness of convolutional neural networks (CNNs) and other deep learning techniques over this dataset [3]. The purpose is to develop the software with a very high accuracy rate and with minimal time and space complexity and also optimal and with the good graphical interface for good interaction between user and the system.*

Keywords: *Handwritten Character Recognition, Convolutional Neural Network, k-nearest neighbor's Algorithm, Tensor Flow, Tesseract OCR, Soft Max Regression, Google API, Neural Network, Artificial intelligence and Machine Learning.*

1. INTRODUCTION

In recent years, in-depth learning-based strategies have been gaining great interest in a research community to solve a variety of supervised, unresolved and reinforced learning problems [1]. One of the most well-known and widely used methods of convolutional neural networks (CNNs), is a type of neural networks that can automatically extract relevant features from input data [7].

Handwriting recognition is the machine's ability to receive and translate handwriting input from multiple sources such as text, images, and touch screen devices. The adoption of handwritten and mechanical characters is a emerging field of research and finds many programs in banks, offices and industries. The main objective of this project is to design a professional program, "HCR using Neural Network" which can detect a specific character of a type of format using the Artificial Neural Network method [2]. By using this approach, we will translate the recognize letter in Devanagari languages such Marathi and Hindi. For translating the word, we will use Encoder-Decoder Architecture for translation English to Marathi language.

This can be helped by using the API facility provided by Google, in this case, is Google Translate which already has various languages in the world until the Marathi/Hindi language as its output [5]. So, this research proposes to implement Firstly to recognize the word then translate in local language with Neural Network approach on web operating system.

2. LITERATURE SURVEY

Sr. No	Title of paper	Publisher by	Publishing year	Overview
1	HCR Using NN and TensorFlow	International Journal of Innovative Technology and Exploring Engineering(IJITEE)	2019	Recognition of digit, Image Vision are some area of study field where machine can read electronic text into digital text
2	HCR Using Neural Network	International Journal Of Advance Research And Innovative Ideas In Education	2009	Effectively recognize all English character of type format using Artificial NN approach.

3	Handwritten Digit Recognition using Machine Learning Algorithms	Global Journal Of Computer Science And Technology	2018	In this paper, there is a MNIST database it is commonly used for training various image processing system.
4	Word Level Eng to Mr NMT using Encoder-Decoder Model	Harshall L.	2019	This paper gives an idea about machine translation and used architecture to build sequence to sequence Model.
5	Javanese Script Translator based on Indonesian speech recognition using Sphinx 4 & Google API	International Seminar on Research of Information Technology & Intelligent System(ISIRTI)	2018	In this paper it can dynamically translate text between many languages pair.
6	Comparative Analysis Of Advanced Classification Techniques For Multilingual Ocr Systems	INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH	2019	By this paper we get the idea about many other Classifier and their Accuracy
7	Analogizing	International	2017	This paper

	Time Complexity of KNN and CNN in Recognizing Handwritten Digits	Conference on Image Information Processing (ICIIP)		provides use with the knowledge of input layer neuron, output neuron & hidden neuron.
8	Optical Character Recognition by Open Source OCR Tool Tesseract A Case Study	International Journal of Computer Applications	2012	This paper gives an idea about different ML library that can be used in OCR.

3. Problem Definition

Handwriting number recognition is a challenging problem. Character recognition and classification are done by Neural Network. . Researchers have been researching the area for a long time, especially over the years. In our study there are many categories related to numbers, for example, bank checks or identifying numbers on car plates, the topic of digital recognition emerges [3]. Much work has been done on various non-Indic scripts in particular, in the case of the Roman, but, in the case of the Indic Scriptures, research is limited. Though translation can be used for removing this Communication Gap and with the use of Internet, Translation software can help-native-speaking individuals understand web pages published in other language [4].

For example, different users had their own writing styles where here the biggest challenge falls to allowing a computer program to understand these different writing styles and see them as standard writing. With the help of machine learning libraries for implementation purpose we use tesseract [8].

Chapter 4
Existing System

4.1 Introduction

This paper aims about the communication gap between the user handwritten style and system recognize. Communication has become a very important and necessary part of our daily lives to accomplish our tasks [1]. HCR is usually divided into six categories namely image acquisition, pre-image processing input, segmentation, feature extraction, segmentation and post editing [2]. The following block diagram of the basic recognition of character is shown in figure 4.1.

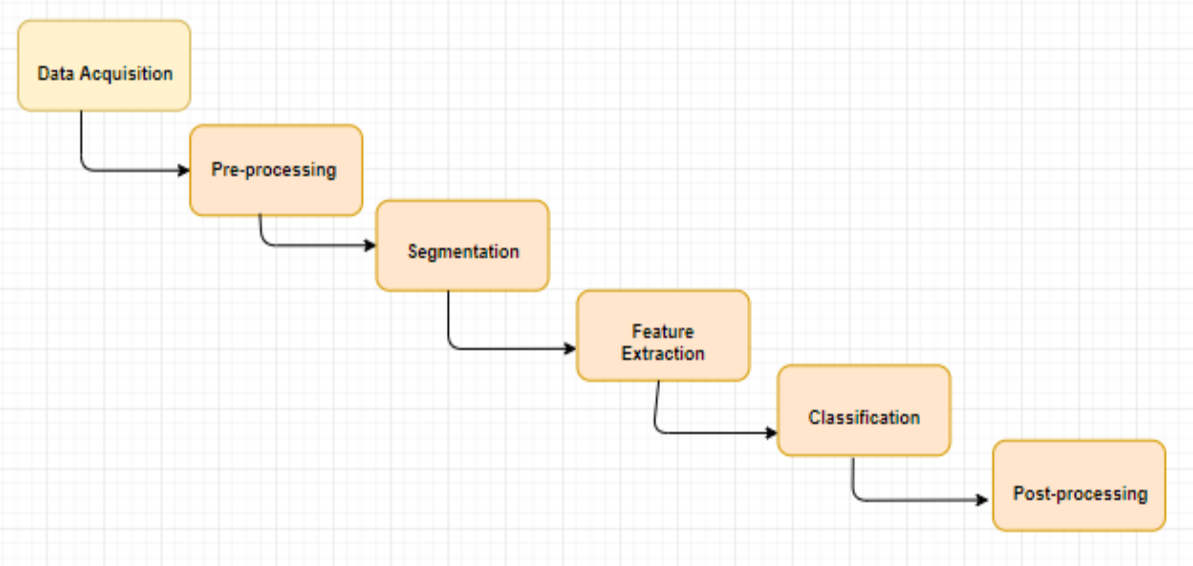

Figure 4.1 Block Diagram of HCR

- Data Acquisition

  In the data acquisition phase the input is taken for recognition of the images. Various type of images is to be give to these phase such as ".jpeg",".png", etc.

- Pre-processing

  The second method called Pre-processing is the entry method for character recognition and is very important in capturing visual acuity. Advance preparation works to turn the lashes and remove variations that can reduce the level of accuracy.

- Segmentation

  Separation is used to convert a multi-character input image into single characters. Strategies used to separate words, line and characters.

- Feature Extraction

  The purpose of the feature removal is to allow for the removal of the most important pattern in the separation. Other Feature Strategies. Such as Principle Component Analysis (PCA), Scale Invariant Feature Extraction (SIFT), Linear Discriminant Analysis (LDA), Histogram, Chain Code (CC), design and Gradient-based features can be used to extract features of individual characters..

- Classification

  Decision-making is done in the classification phase. By seeing the characters, extracted features are used. Different classifiers like SVM and Neural Networks are used.

- Post-Processing

The final and last phase of the character recognition is the Post-processing. It is the procedure for correcting the misclassified output by using natural language. It processes output by getting it after the shape have been recognized.

4.2 Disadvantages

- The final image quality depends on the quality of the original image.
- The images lose some quality during scanning and digital processing.

5. Proposed System

In the proposed system, we will understanding the handwritten characters or typed documents is simple to the human beings as we have the ability to learn. The same ability can be induced to the Machines also by the use of Machine Learning and Artificial Intelligence. The field which deals with this problem is called as the OCR or also known as Optical Character Recognition.

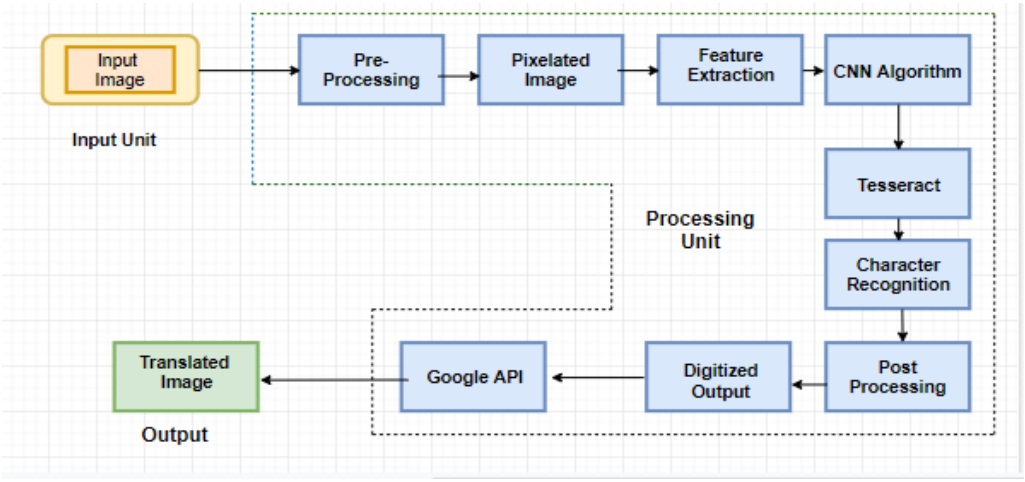

Figure 5.1: Architecture Of Proposed System

In this diagram it represent the architecture of the proposed system. In the proposed system it contains the 3-Phase , Input , Processing Unit and Output. In the Input phase, input is given to it and then it goes to next phase that is Processing Unit.In this unit the all processing is done with the help of CNN Algorithm and Tesseract. Then character recognition is done and the output generated , the last phase gives the output of the image and Finally this image is translated with the help of the Google API Translator.

5.1 KNN Classifier

KNN is an important classification algorithm which has countless applications in Machine Learning (ML). This algorithm is based on the approach of classifying the objects on the basis of closest training match in the feature domain. Using the approach of k-NN an object is classified on the basis of maximum votes of its k neighbors and i\

f k = 1, only class is assigned to that object [6]. Though KNN can be used for both classification and regression predictive problems but high degree of accuracy and desirable results are achieved in classification applications.

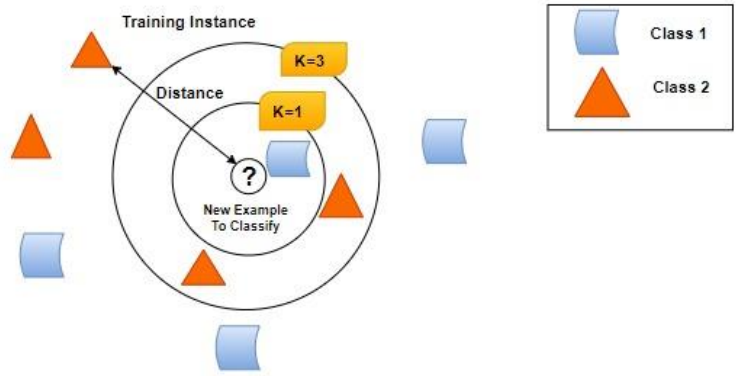

Figure 5.2: KNN Classification Diagram

5.2 CNN Classifier

A convolutional neural network is a FFN (feed-forward neural) network that is generally used to analyze visual images by processing data with grid-like topology. It's also called as a ConvNet [7]. A convolutional neural network is used to find and separate objects in an image. While on CNN, the weighted convolutional layer used for feature extraction and the fully connected layer used for separation is determined during the training phase.

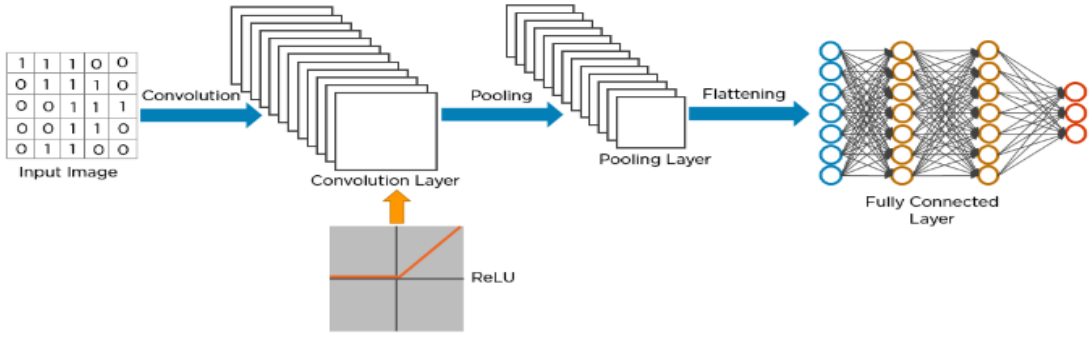

Figure 5.3: CNN Classification Diagram

5.3 Analysis between CNN & KNN

KNN algorithm having useful case for recognize character through feature extraction method. For that we can find the root sum of pixel with wide pixel. For above process we required Eucidian Distance method. For KNN we use Gaussian distance method for recognize we get the accuracy about 96.20%.Where is for CNN it has an combination of three layers they are

maxpooling, classification & convolution. MNIST dataset has been taken for CNN in that feature has been send as a input common known as Neurons. Neuron has been calculate in pixels. From that we get the output image known as “Feature Map”. The main idea is to reduce the value of pixel for that we can use downsampling method. This result pass to pooling layer which take 2x2 region as kernel. The layer which has 10 neuron are called as classification layer. By this we get accuracy of CNN algorithm is 98.10% [7].

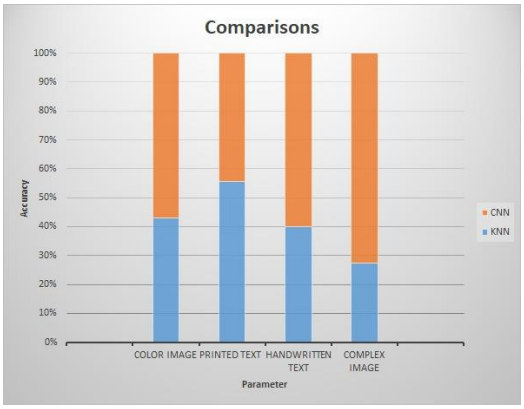

Figure 5.4: Comparison between CNN & KNN

5.4 Tesseract

Tesseract is an open source OCR engine. It is more portable. It has been focused towards very less rejection than accuracy. Tesseract can work with different languages. It is suitable for backend & can be used for difficult OCR task can be done using in frontend. Tesseract execute from command line interface [8].

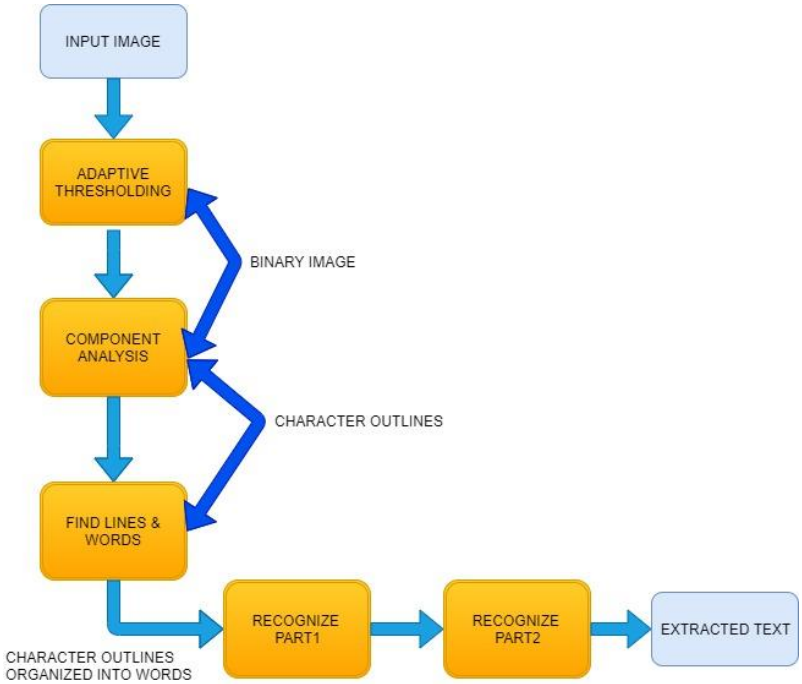

Figure 5.5: Tesseract Architecture

As our main aim to recognize whole paragraph and then that paragraph has to be translate into Marathi language. For that we have taken tesseract machine learning library & Convolution Neural Network as it make quickly & its error rate is also less than other.

6. Result

1.First the input will be taken through image.

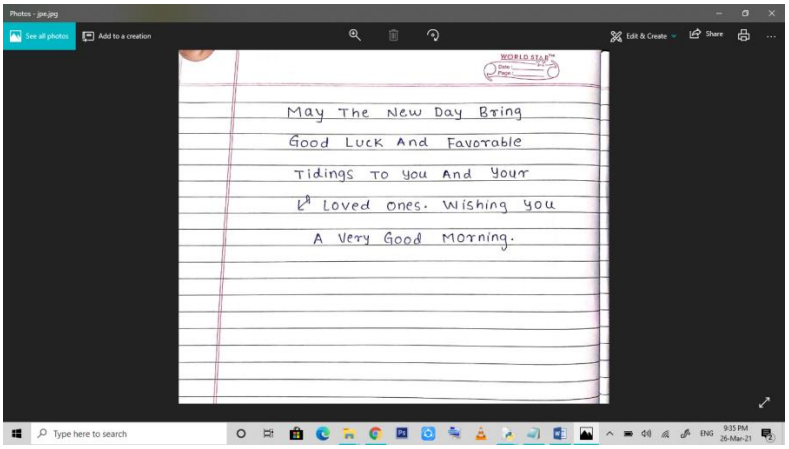

Figure 6.1: Input Image

2. Through Tesseract it will recognized the text and display the expected output on the same image screen.

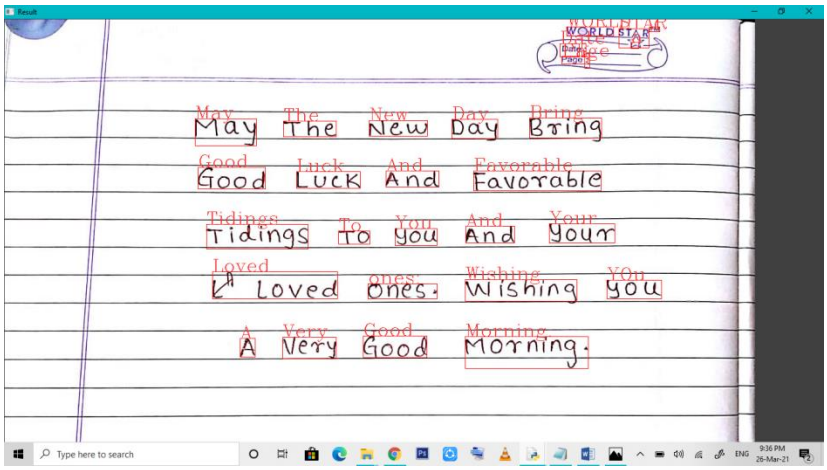

Figure 6.2: Recognizing Input Image

3. In below image, the new Notepad file is created in which the translated output is saved and displayed.

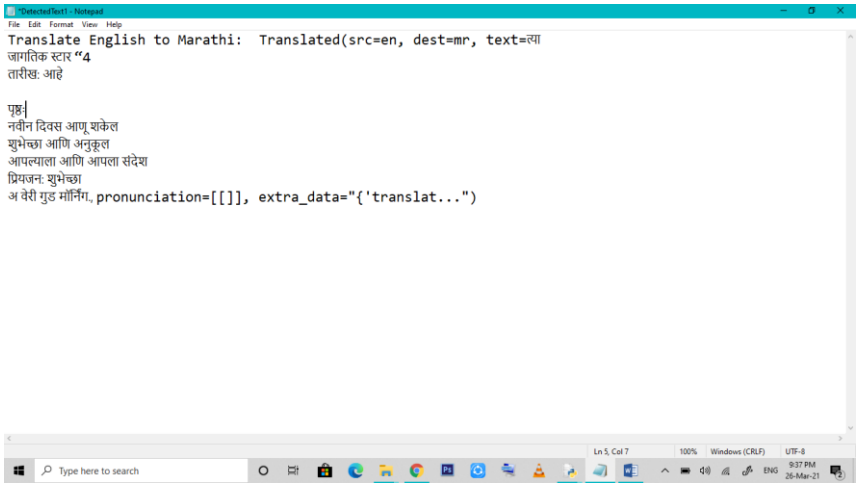

Figure 6.3: Translation of Input Image

7. Conclusion

Many regional languages around the world have a variety of writing styles that can be identified by HCR programs using the appropriate algorithm and techniques. We had already learned a lesson related to the acceptance of English letters. It has been found that the recognition of a handwritten character is difficult due to the presence of unfamiliar characters or similarities in the characters of many characters. The scanned image is processed first to get the image cleared and the characters separated by individual characters.

Progress work is done when performing normal, filtration is done using processing steps that produce sound output and cleanliness. The use of other mathematical and geometric features by the neural network will provide a better visual effect of English letters. This work will help future researchers.

8. Future Scope

The scope of the future can be increased significantly by making it easier to use. Many programs are available where you may wish to read handwritten input. Reading handwritten is a very difficult task considering the diversities that exist in ordinary penmanship. However, progress is being made.

9. Acknowledgement

We sincerely wish to thank our Project guide Prof **. Reena Deshmukh** for her ever encouraging and inspiring guidance helped us to make our project a success. Our project guide made us ensure with her expert guidance, kind advice and timely motivation which helped us to determine about our project.

We would like to thank our Project Coordinator **Dr Uttara Gogate** for all the support we needed from our project.

We also express our deepest thanks to our H.O.D. **Dr. P.R. Rodge** whose kindness helps us making available the computer facilities to us for our project in our laboratory. Without his kind and keen co-operation our project would have been stifled to stand still.

Finally, we would like to thank our college principal **Dr. J.W. Bakal** provides laboratory services and allows us to continue our work. We would also like to thank our colleagues who helped us directly or indirectly during our project.

REFERENCES

[1]. M. Agarwal, Shalika, V. Tomar, Priyanka G., "Writing Status Using Neural Network and Tensor Flow JITE ISSN: 2278-3075, Volume 8, Magazine 6S4, April 2019

[2]. Hitesh Mohapatra Veer Surendra Sai University of Technology," Handwritten Character Recognition (HCR)Using Neural Network" Thesis · September 2009

[3]. S M Shamim, Mohammad B. Alam Miah, A. Sarker, M. Rana & Abdullah Al Jobair, "HDR using ML Algorithms", Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence Volume 18 Issue 1 Version 1.0 2018

[4]. Siddhartha Roy, M.Saravanan Word Level English to Marathi Neural Machine Translation using Encoder-Decoder Model Vol-3 Issue-5 2017

[5]. A. Nursetyo, De R. Ignatius Moses ," Lat Aks Late: Javanese Script Translator based on Indonesian Speech Recognition using Sphinx-4 and Google API", 2018 IEEE

[6]. Rohit Verma, Dr. Jahid Ali, "Comparative Analysis Of Advanced Classification Techniques For Multilingual Ocr Systems" INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 11, NOVEMBER 2019 ISSN 2277-8616

[7]. Tanya Makkar, Yogesh Kumar, Ashwami Kr Dubey, Alvaro Rocha, Ayush Goyal, "Analogizing Time Complexity of KNN and CNN in Recognizing Handwritten Digits", Fourth International Conference on Photographic Data Processing (ICIIP)

[8]. C. Patel, A. Patel, D. Patel, "OCR by Open Source OCR Tool Tesseract: A Case Study" International Journal of Computer Applications (0975 - 8887) Volume 55– No. 10, October 2012