**SYNOPSIS REPORT**

**OPTICAL CHARACTER RECOGNITION FOR THE TEXTURAL DATA**

**Submitted by**

Yash Jangid (Enroll No. R970218068)

Rishabh Mishra (Enroll. No. R970218044)

Abhinav Kamboj(Enroll. No. R970218067)

Arjun Gupta(Enroll No. R970218010)

BRANCH :- Oil & Gas Informatics

SEMESTER :- VI

**Under the guidance of**

Mr. Rahul Kumar Singh

Assistant Professor

Department of Informatics



**SCHOOL OF COMPUTER SCIENCE**

UNIVERSITY OF PETROLEUM & ENERGY STUDIES

Bidholi Campus, Energy Acres, Dehradun – 248007.

**Project Proposal Approval Form (2021)**

**Minor-2**

**PROJECT TITLE:** OPTICAL CHARACTER RECOGNITION FOR THE TEXTURAL DATA

**ABSTRACT-**

In this minor project our group have to present a character recognition system for getting text and numbers from the input image data with the help of python script and algorithm named as KNN (k nearest neighbor) algorithm. This project will convert scanned image into binary numbers and extended features passes from KNN classifier. The characters are recognized by these binary numbers by machine by using python language and no predefined library are used during the whole project, like tesseract and easy OCR.

**KEYWORDS:**

KNN, Feature extraction, scanning etc.

**INTRODUCTION**

Optical Character Recognition (OCR), is the electronic conversion of images or printed text into machine encoded text.

In running world there is huge demand for the software system to recognize character in computer system. As we know that these days there is a huge demand for storing the information available in these paper documents in a computer storage disk. For this problem one simple way to store information in these paper documents in to a computer system is to first scan the document then store as a image but to reuse this information it is very difficult to read the individual content and search the content.

OCR is the process of taking an image of letters or typed text and converting it into data that computer understands. The exact mechanism that allows humans to capture an image of a paper document after which the text is extracted from that image.

**OBJECTIVES-**

* Recognize text and number from the input image.
* To speed up the recognition.

**METHODOLOGY**

In order to complete this project, we use KNN (K nearest neighbors) algorithm, by taking input as an image from MNSIT data set in a format of .gz file then, after copy the image data into a directory then by linking the directory we perform the main task of extraction.

The methodology consists of two steps: The first two steps refer to creating a database for training using a set of documents, while the second one refers to recognition of new document images. First, a pre-processing step that converts image into binary and enhancement takes place. At a second step a approach is used in order to detect text lines, words and characters

And for this we use a KNN algorithm with machine learning approach.

First we have to unzip all the image data then convert them into the frame of 28\*28 in order to improve the accuracy. Then perform a future extraction with the help of knn algorithm.

And also working on to increase the accuracy because as compare to the predefined library the time complexity is high, but it shows 100 percent accuracy on the data which we have taken from the mnsit data, also it gives 85-90 percent accuracy on other format data.

**SYSTEM REQUIREMENTS**

1. Hardware:

* RAM: 4GB

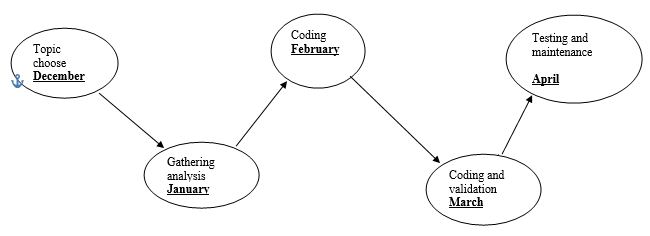
2.) Software:

* Any python-compiler

3.) Operating System:

* Windows or Linux

**SCHEDULE**-



**PERT CHART**

**REFERENCES:-**

**http://yann.lecun.com/exdb/mnist/**[**https://www.sciencedirect.com/topics/computer-science/optical-character-recognition**](https://www.sciencedirect.com/topics/computer-science/optical-character-recognition)[**https://github.com/yash500071155/Optical-character-recognition-for-the-textural-data/blob/main/data**](https://github.com/yash500071155/Optical-character-recognition-for-the-textural-data/blob/main/data) **https://nanonets.com/blog/ocr-with-tesseract/**

**Synopsis Draft verified by**

**Project Guide HOD**

**(Name & Sign) (Dept. of Systemics)**

**INSTRUCTIONS**

1. Min/Max. no. pages: **5** pages excluding Front Page.
2. The contents in the document should be provided in such a way that the **Verification of the Synopsis Report** should be available in the last page along with the contents and NOT in a separate page.
3. Type your content in this document itself with all Pages except the front page numbered. Bottom, Center, Times New Roman, Bold, Font size 10.
4. Margins: Left(1”), Right (0.75), Top & Bottom(0.6”).
5. Chapter Heading: Center Aligned, Bold, All Caps, Time New Roman, Font size 16
6. Main Heading: Heading under a chapter should be Left Aligned, Bold, Not underlined & not italicized, Times New Roman, Font size 14
7. Sub-Heading: Left Aligned, Bold, Not underlined & not italicized, Times New Roman, Font size 12
8. Content Format: Justified, Not Bold & Underlined, Not Italicized, Time New Roman, Font size: 12, line spacing: 1.5