Abstract— This paper presents a new and easy approach for extracting the text from image. The main aim of the project is to convert the old printed text into the digitized form so as to increase the durability of the text and reduce the maintenance cost. There are many different approaches to convert the printed text to digitized text but many of these methods are tedious for the user to use. So in this paper we are trying to give easier approach using matrix matching concept, by implementing the project we are able to extract the text in user friendly and efficient manner.

Keywords— Categorization, Feature Extraction, OCR, Pre-Processing, Segmentation.

I. INTRODUCTION

Image processing is a very diverse field of technology. Human are very eager to train their system to understand the ancient human readable format to machine readable format. Since many decades researchers are working on this aim and through their rigorous effort they have discovered OCR as a very successful application to convert human readable text to machine readable.

The techniques like neural networks, structural and statistical pattern are available for recognition of text. But the major drawback of neural network is large training data which takes much more time to build this makes the neural network more complicated for a naive user to understand leading it to less user friendly.

In this paper we are presenting a way for extraction of text from image which is simple as well as user friendly. We are also introducing the concept of categorization of text in image. We have used the concept of matrix matching for recognizing the text.

II. LITERATURE REVIEW

OCR technology is used for text extraction. Handwriting recognition is classified into two types as off-line and online recognition method. The work is carried out using LABVIEW. For recognition, pattern matching algorithm is used and segmentation method is used with neural network. The recognition rate of neural network is 99.9% for the handwritten word and for cursive handwriting with the accuracy of 70-80%. [1]

All type of cursive handwriting can convert into digital text using an IWR technique. An IWR recognizes only meaningful words, actually it checks semantic. For separating the cursive and tightly coupled handwriting segmentation technique is used. To locate the prospective segmentation points in handwritten words a heuristic feature detection algorithm is used. Segmentation algorithm recognises word as pre-process image is extracted from database and then Extracted texts were converted and analyzed using Intelligent Word Recognition technique. [2]

Another new technique for offline handwritten text is Gurmukhi character recognition. For classification, three different classifiers are used i.e. SVM, MLP and Naïve bayes, with 5 fold and 10 fold cross validations are used to recognize various characters. Here 91.95% accuracy is achieved with SVM (Polynomial Kernel) with 10 fold cross validation and 87.30% accuracy is achieved with MLP with 10 fold cross validation and 77.70% accuracy is achieved with naïve bayes with 5 fold cross validation.[3]

Hidden Markov Model (HMM) for recognition of cursive handwritten English characters. By applying median filter it avoids errors caused by noise in the scanned image. To reduce the complexities in the recognition process high quality samples are used. Here better results are obtained in terms of accuracy as well as speed. [4]

The FPGA based approach known as sobel edge detection is a parallel architecture which reduces the complexity of design and also reduces the processing time. Noise produces various spurious edges in the image which can be reduce by using the sobel operator which gives a smoothing effect. It shows good performance with maximum frequency of 200MHz of edge detection.[5]

III. METHODOLOGY

In this paper we have created a data set for the matching purpose which is consisting of English alphabets both in lower case and upper case (a-z, A-Z) as well as the numeric (0-9). For each character in data set we have created a matrix which is further use in recognition step.
This paper consists of 5 steps. They are image acquisition, image pre-processing, segmentation of character, feature extraction of characters, character recognition and text categorization.

A. Image Acquisition
The input image is acquired by digital professional camera or scanned copy of an image.

B. Image Pre-processing
The input is not suitable for recognition process. Hence, the input image needs to undergo a pre-processing step to convert into usable for further steps. For pre-processing we use Binarization technique. It is procedure to convert coloured or gray scale image into black and white. Pre-processing the input image is very important to make it more accurate for further steps.

C. Segmentation
Segmentation step is one of the most important and difficult task in text extraction and recognition. Segmentation is the process of decomposition of different objects by extracting their respective boundaries and the text component is isolated from the background.

In this step the input pre-processed image consisting of sequence of character is there by decomposed into sub-images. Firstly the image is segmented line by line, then the line is segmented word by word and further the word is segmented into characters. The segmented character is provided as an input for the next step.

D. Feature Extraction
Feature extraction is a crucial step for recognition of text. It is the process extracting differentiating features of digitized characters from their respective matrices. For each input segmented character we create a matrix and then this matrix is matched with the already stored matrix of characters in the data set.

E. Recognition
In the recognition step the matrix formed in feature extraction for segmented character are mated with stored matrix of character in data set. If the matrix is matched then the character is recognized and printed in word file.

IV. CATEGORIZATION
In this paper we are also trying to present the new concept for the categorization of the text which will make the user to understand the text context and will make that text more user friendly in less effort.

We are categorizing the text as:-

i. Interrogative
ii. Exclamatory
iii. Compound statement

V. IMPLEMENTATION
While implementing our idea we have created a GUI using MATLAB.

While running the program the user is able to select the particular scanned image form system by pressing load image button.
While we need to extract the text from the image we need to reduce the noise in the image which can be done by pressing gray image button.

After pre-processing the image we need to recognize the particular character which is segmented by matching the matrix obtained after all processing on segmented character with the already stored dataset in the form of matrices. The user will be able to receive the output in upper case, lower case as well as it is present in image. User can choose the option as per various buttons provided in the GUI which can be any one of upper case, lower case or normal case. The text will get extracted in the word document.

In this paper we have tried to categorize the text in the image as per context required. The user can select from the various buttons like interrogative, exclamatory, compound statement. The user will receive the output in the highlighted form.
VI. APPLICATIONS

This kind of application is used in the various fields where there is need to store large printed text. The application will make the data more manageable and more durable. Once the data is stored in digitized format the promotion of the text and its validation become easy. By categorizing the text the context of the text can easily be understood. The application is helpful in the field of education as it can make easier for the students and teachers make knowledge more interactive and attractive. Also the application can be use in preserving the old books, documents in the libraries more sustainable.

VII. CONCLUSION

In this paper we have successfully extracted the text from the scanned image and also categorized the text. It is necessary or the application that the image must be scanned so as to reduce the noise in the capturing process. The application is efficiently working on the non-connected letters and categorizing it.

VIII. FUTURE SCOPE

In the future we can try to implement the extraction on the connected, tightly bound component. The translation of the text can also be implemented in future. There is a scope of recognizing the characters with different fonts and style.

REFERENCES