



**An Intelligent User Based Text Recognition Using Analytical  
Approach**

A Dissertation

Submitted

By

**VINIT KUMAR**

**11301270**

To

**Department of Computer Science & Engineering**

In Fulfillment of the Requirement for the

Award of the Degree of

**Master of Technology in Computer Science Engineering**

Under the Guidance of

**Mr. Harsh Bansal**

**Assistant Professor**

**(May 2015)**



School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student : VINIT KUMAR  
Batch : 2013-2015  
Session : 2014-2015

Registration No : 11301270  
Roll No : RK2307A14  
Parent Section : K2307

Details of Supervisor:

Name : HARSH BANSAL  
UID : 16866

Designation : Assistant Professor  
Qualification : M.TECH  
Research Exp. : 2 years

Specialization Area: Database (pick from list of provided specialization areas by DAA)

Proposed Topics:-

1. EFFICIENT TEXT RECOGNITION USING HOLISTIC APPROACH
2. HANDWRITTEN WORD RECOGNITION IN COMPLEX DOCUMENTS
3. CHARACTER RECOGNITION USING GRADIENT METHOD

*Harsh Bansal*  
16866  
Signature of supervisor

PAC Remarks:

*first topic approved*  
*klh*  
*11/7/14*

Signature: *[Signature]*  
11/01/14

Date:

APPROVAL OF PAC CHAIRMAN

- \*Supervision should finally encircle one topic out of three proposed topics and put up for an approval before Project Approval Committee (PAC).
- \*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.
- \*One copy to be submitted to supervisor.

## **ABSTRACT**

In the present work we represent an intelligent user based text recognition using analytical approach. It's an offline handwritten words recognition method. In present work we perform recognition on the English words which is collected from multiple people in different handwriting style. We use analytical approach for recognition to the words. Analytical approach recognize to the word character by character. We uses neural network with back propagation approach for recognition word. Because we working with an analytical approach so neural network recognize the word character by character. If some character is not recognized perfectly than we uses user based interface. In user based interface, the user gives an input by the keyboard to character which is not recognized by neural network. We achieved average accuracy 89 % without user based interface and the accuracy will be increase 8% to 9% with the user based interface.

**KEYWORDS:** Analytical Approach, Neural Network, User based interface

## **ACKNOWLEDGEMENT**

First of all, I would like to thank my Almighty God, who has always blessed me and for giving me strength to do this work.

I wish to express my deep gratitude to my guide, Mr. Harsh Bansal, for his generous guidance. His guidance and support throughout all stages of the thesis process enabled me to conduct the research. Without his support, I would not be possible to complete this program.

Then I would like to offer my sincerest gratitude to Mr. Dalwinder Singh, Head of the Department, Computer Science & Engineering, for providing all the facilities and environment.

Vinit Kumar

## DECLARATION

I hereby declare that the dissertation entitled “**An Intelligent User Based Text Recognition Using Analytical Approach**” submission for the M.Tech degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:

Investigator

Vinit kumar

Reg. No. 11301270

## CERTIFICATION

This is to certify that Vinit Kumar has prorating Master of technology (M.Tech) dissertation “**An Intelligent User Based Text Recognition Using Analytical Approach**” under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engineering.

Date: 29 November 2014

Signature of Advisor

Name: Harsh Bansal

(Assistant Prof.)

UID: 16866

Lovely Professional University

# TABLE OF CONTENT

<b>Introduction.....</b>	<b>1</b>
1.1 Fields of NLP .....	1
1.2 Applications of Natural Language Processing.....	2
1.3 Optical Character Recognition (OCR) .....	3
1.4 Optical Character Recognition need .....	3
1.5 Steps in OCR .....	4
1.5.1 Data collection.....	4
1.5.2 Pre-Processing.....	4
1.5.3 Segmentation.....	4
1.5.4 Feature Extraction.....	4
1.5.5 Classification.....	4
1.5.6 Post-Processing.....	5
1.6 Problem in character recognition .....	5
1.7 Handwritten recognition .....	5
1.7.1 Online handwritten recognition .....	6
1.7.2 Offline handwritten recognition .....	7
1.7.3 Technique used in handwritten recognition .....	7
<b>Review of Literature.....</b>	<b>17</b>
<b>Present work.....</b>	<b>26</b>
3.1 Problem formulation.....	26
3.2 Objective.....	27
3.3 Research methodology .....	28
<b>Results and discussions.....</b>	<b>31</b>
<b>Conclusion &amp; future work.....</b>	<b>37</b>
<b>References .....</b>	<b>38</b>
<b>Appendix .....</b>	<b>39</b>

## LIST OF FIGURES

<b>Figure 1.1</b>	The different area of character recognition.....	<b>3</b>
<b>Figure 1.2</b>	OCR Process.....	<b>5</b>
<b>Figure 1.3</b>	Rules for handwritten recognition.....	<b>7</b>
<b>Figure 1.4</b>	Segmentation Algorithm Graph.....	<b>10</b>
<b>Figure 1.5</b>	Fourier Descriptors.....	<b>11</b>
<b>Figure 3.1</b>	Flow chart of purposed work .....	<b>25</b>
<b>Figure 4.1</b>	Handwritten set in upper case.....	<b>28</b>
<b>Figure 4.2</b>	Handwritten set in upper and lower case.....	<b>29</b>
<b>Figure 4.3</b>	Handwritten set in lower case .....	<b>29</b>
<b>Figure 4.4</b>	Neural network tanning data set .....	<b>30</b>
<b>Figure 4.5</b>	Input handwritten text image .....	<b>30</b>
<b>Figure 4.6</b>	Binarized image with noise .....	<b>31</b>
<b>Figure 4.7</b>	Remove noise from image .....	<b>31</b>
<b>Figure 4.8</b>	Text segmentation with matlab code .....	<b>32</b>
<b>Figure 4.9</b>	Extract the character and words for segmented image .....	<b>32</b>
<b>Figure 4.10</b>	Recognition interface .....	<b>33</b>
<b>Figure 4.11</b>	Recognition with neural network .....	<b>33</b>
<b>Figure 4.12</b>	Display those characters which is not recognized .....	<b>34</b>
<b>Figure 4.13</b>	User based input character ‘e’ which is not recognized and display result..	<b>35</b>
<b>Figure 4.14</b>	User based input character ‘b’ which is not recognized and display result..	<b>35</b>
<b>Figure 4.15</b>	Display the final recognition text file with neural recognition file .....	<b>36</b>





# CHAPTER 1

## INTRODUCTION

---

Natural language Processing (NLP) is a branch of computer science where we focus on developing a system that is used for communication with people using everyday language. It is also called computational linguistics because it uses computational method for understanding human language. It is a process in which human makes communication with machine easily. NLP is act as an interface for human and computer interaction. It reduced distance between human and machine because it is a process where machine works like a human. There are many application developed last few years. A very useful application is that a machine takes instruction by human voice and follows operation on it. NLP are trying to make computer more reliable that are easier to be uses by people. So rather than learning a special language of computer command, people will talk with computer in their own language.

### 1.1 Fields of NLP

- i. **Automatic Summarization:** It works as the summary of a chunk of text. Like, article in political section in newspaper.
- ii. **Machine translation:** It is very important and mostly useful task of nlp. In machine translation process system translate automatically language from one human to another human.
- iii. **Optical Character Recognition:** OCR is a process mechanical or electronic conversion of scanned or photographic image of typewritten or printed text into machine encode or computer readable text. Optical character recognition, recognize printed and handwritten documents with the help of different models.
- iv. **Speech recognition:** In speech recognition takes a sound clip of the person and determines the textual representation of speech.
- v. **Speech segmentation:** In speech segmentation take a sound clip on the person and separate into words. It is a challenging task in NLP.

- vi. **Word segmentation:** Word segmentation takes the chunk of continuous text and overlaps text and then separates this text into words. There are many algorithms that are used for word segmentation.

## **1.2 Applications of Natural Language Processing**

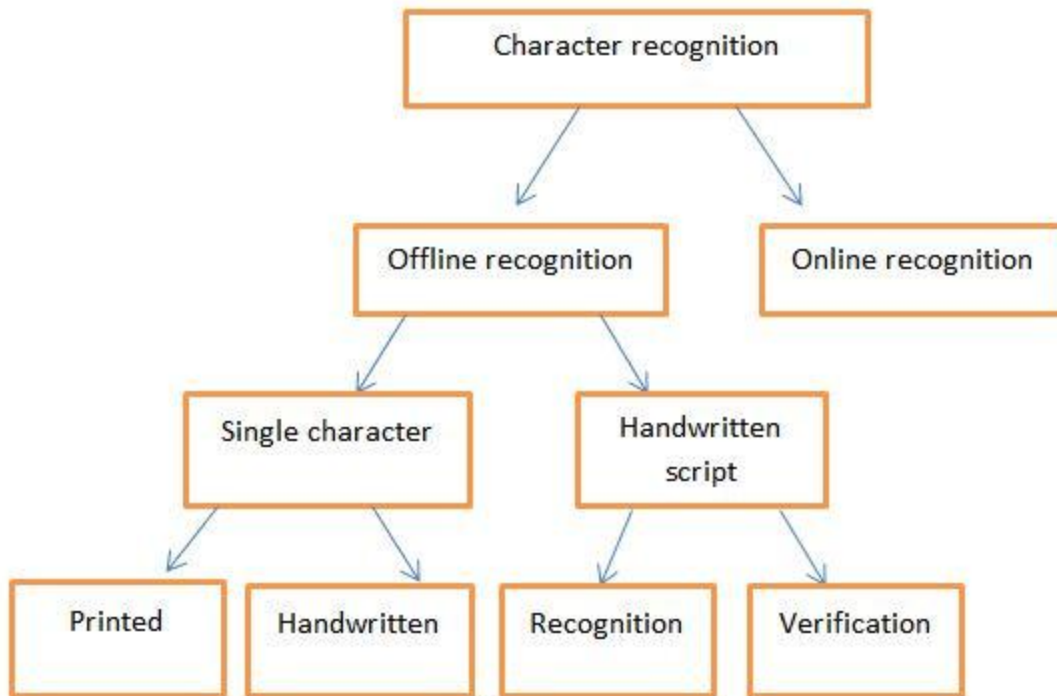
There are large amount of data that is present on the internet approximately 19 billion pages. There is need of NLP for processing huge amount of textual data. Some requirements are:

- Indexing and searching large text
- Automatic summarization, like summarizing complete book into just a few pages or in a single page.
- Question answering
- Speech understanding: Understanding mobile and phone conversation
- Information extraction: Extracting useful information from resumes
- Automatic translation

## **1.3 Optical Character Recognition (OCR)**

OCR is the system that is used for translating images of handwritten, typewritten, and printed text into a format which is easily understood by machines for the purpose of editing, indexing, searching and reduction in storage size. Optical character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. Many commercial systems for performing OCR exist for a variety of applications, although the machines are still not able to compete with human reading capabilities.

There is large amount of data available in the library; information center museums and office create a demand for character recognition. Optical Character Recognition deals with the problem of recognizing optically processed characters. It is used in many commercial or non-commercial works like, data entry from printed paper or handwritten paper, passport documentation, business card, mailing, bank statement, computerized scripts etc. It is very common method for digitalize printed/handwritten data. Basically OCR work on two types of data printed and handwritten data.



**FIGURE 1.1 THE DIFFRENT AREA OF CHARACTER RECOGNITION**

### **1.4 Optical Character Recognition need**

Character recognition is needed when the information should not be readable both to humans and to optical a machine. The basic need of OCR is to recognize the data and used it in an useful aspects. There are few fields in which OCR can play an important role and those are:

- Recognition data entry for business documents like, check, bank statement, passport and receipt.
- Automatic number plate recognition.
- Automatic extraction important information from insurance document.
- Develop technology for blind users.
- Used for extraction business information from contact list.
- Used for converting handwritten document images into digitalize format.

## 1.5 Steps in OCR

**1.5.1 Data collection:** Data which is used in OCR may be printed or handwritten. The source of data collection may be different. In printed, the data should be in printed scripts and in handwritten, data will be handwritten format.

**1.5.2 Pre-Processing:** In preprocessing some operation have to be performed. These are,

- a. Scan the document from scanner
- b. Binarization
- c. Image normalization (remove slant, skew and paper noise)

So the Preprocessing work is to organize the information for become simple it to the character recognition task.

**1.5.3 Segmentation:** Segmentation is an important stage in character recognition because the rate of recognition is depends on segmentation. Segmentation has two categories **external and internal segmentation**. External segmentation is the part of different writing unit such as paragraph, sentence or words. Internal segmentation breaks down sequence of image characters into sub-image of individual character.

**1.5.4 Feature Extraction:** In feature extraction methods selects and prepare the data which is used by the classifier to achieve the recognition task.

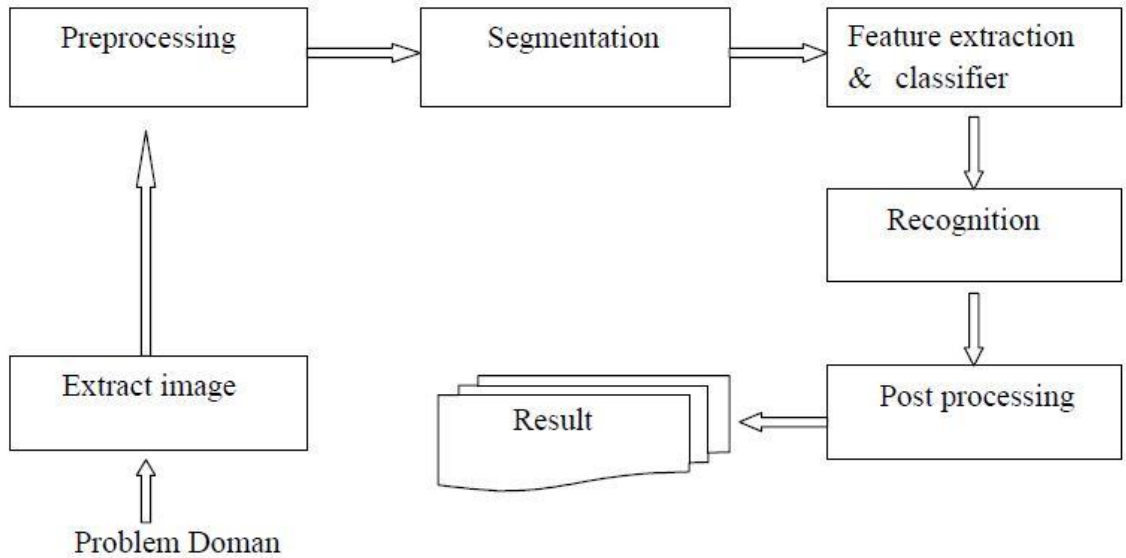
**1.5.5 Classification:** The classification is the process of identifying each character and assigning to it the correct character class. Performance of classifier will be depending on its features.

**1.5.6 Post-processing:** Post processing method is the last method of optical character recognition. It prints the corresponding recognized characters in the structured text form.

## 1.6 Problem in Character Recognition

- i. Variation of the same character due to change of fonts.
- ii. Noise pixels due to scanning of image.
- iii. Incomplete character and handwriting problem

Mixture of text and graphics



**FIGURE1.2 OCR PROCESS**

## 1.7 Handwritten Recognition

Handwritten recognition is a process of recognition, in which the system deal with the handwritten text as input from source. In handwritten recognition input should be in image format. The image of handwritten text may be scanned by scanner. Scanner play an important role in HWR(handwritten recognition) because if scanning of document have not good quality than it may be cause of bad recognition result. In HWR, it consist multiple step before recognition. Those are:

- Scanning
- Preprocessing
- Segmentation
- Recognition
- Post processing
- Output

It is important to know that which types of recognition we are going perform. There are two approaches:

**1.7.1 Online handwritten recognition:** In on-line recognition system use the digitizer which captures directly writing with the order of stroke, speed, pen-up and pen-down information. It means the system recognizes the characters as they are drawn.

The elements of an on-line handwriting recognition interface typically include:

- A pen or stylus for the user to write with.
- A touch sensitive surface, which may be integrated with, or adjacent to, an output display.
- A software application which interprets the movements of the stylus across the writing surface, translating the resulting strokes into digital text.

**1.7.2 Offline handwritten recognition:** Offline recognition uses the data from paper through digital scanner and cameras. The historical handwritten and printed data are comes under offline data and its used as database. It deals with the recognition words after it was written. Offline handwritten recognition is difficult in comparison of online because people have different written style.

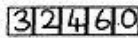

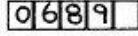
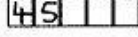

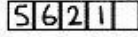
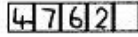
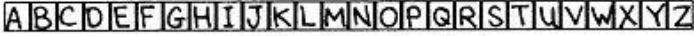
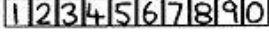
OCR COMPLETION GUIDANCE	
RULES	EXAMPLES
1. Use black pen whenever possible.	Correct:  Incorrect: 3 2 4 6 0
2. Form large characters, but within the box edges.	Correct:  Incorrect: 2 3 7 0 5
3. Use simple shapes, avoid loops or curls or flourishes.	Correct:  Incorrect: 0 6 8 9
4. Close loops.	Correct:  Incorrect: 4 5
5. Connect lines.	Correct:  Incorrect: 4 7 1
6. Do not use alternative shape four continental seven continental one.	Correct:  Incorrect: 5 6 2 1
7. Do not link characters.	Correct:  Incorrect: 4 7 6 2
8. Do not overlap characters.	
Alpha Character Set	
Numeric Character Set	

FIGURE 1.3 RULES FOR HANDWRITTEN RECOGNITION

### 1.7.3 Technique used in handwritten recognition

#### 1.7.3.1 Data Collection

Data collection sources may be different, depending upon the task. If we are going to performing recognition on Bengali handwritten document then we use CMATERdb1.1.1 and if we want perform recognition on mixture document of Bengali with English then use the CAMTERdb1.2.1 dataset. ICDAR contain the dataset in English, French, German and Greek. IAMonDB contain the database of handwritten in English acquired from the whiteboard. It is used for online character recognition. The most popular databases in this field are NIST, CENPARMI, and CEDAR etc. These are the permanent source of data collection. We can collect the data from survey and other resources.



### **1.7.3.2 Scan Document**

In this step document scanned which are going for recognition. The scanner must have good quality. Scan document collect in the image form, but the scanned image may be containing some ambiguity which create problem during recognition. In any type of recognition scanners are used, which have a transport mechanism plus a sensing device which converts light intensity into gray-levels. Handwritten documents normally have different color on a white background. Hence, when performing recognition, it is common to convert the multicolor image into a black and white image. Often this process is known as thresholding. Threshold is work on the scanner to save memory space and mathematically calculation. Thresholding is important because the quality of scanned image will be effect on the recognition result. Still, the thresholding performed on the scanner is usually very simple. Perform threshold on scanner is quite simple, fixed threshold is used and below values is considered as the black whereas above values is considered as the white.

### **1.7.3.3Preprocessing**

Pre-processing is used for enhance the quality of scanned image. It work on image as,

- i. Remove noise: when we give an image as an input it has many unwanted dots or spot which create problem in further processing. There are many methods for removing noise from image through image processing like, median filter method, threshold technique for remove paper and slant noise.
- ii. De-skew: Because people have different handwriting style. If person is writing on plain paper than it is cause of skew. It may be need rotate a few degree clockwise or anticlockwise for removing skew and make the data perfectly horizontal or vertically. Hough transformation is commonly used for detecting skew.
- iii. Binarization: Binarization is technique in which an image is converted into gray image (black and white). The black and white pixel have value in the 0(zero) and 1(one) form.
- iv. Smoothing: The smoothing is using for filling and thinning. During the binirazation some pixel values is week because of pen tip and missing some pixels from there. So smoothing is used for filling the missed pixels.

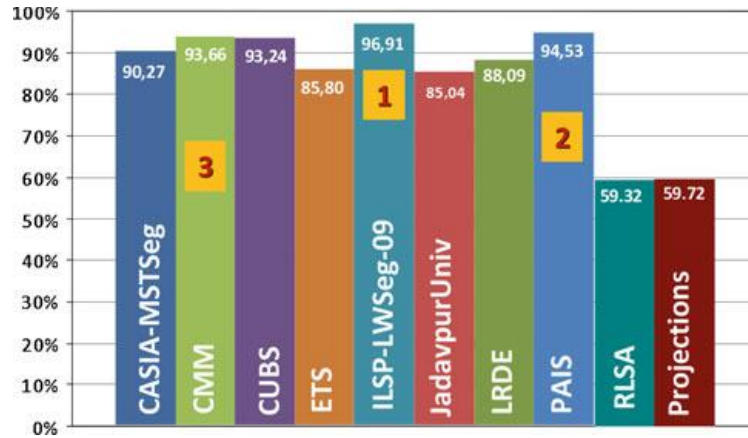
- v. Line remove: In this step we remove all the non-graph boxes and lines which may be create problem in recognition.

#### **1.7.3.4 Segmentation**

Now after the preprocessing segmentation will be apply over the handwritten text image. As the name show it segment the complete text image into lines than lines segment into words and at last words segments into character.

- Line segmentation, there are many algorithm which is used for line segmentation for handwritten script. The basic approach is connected component for line extraction.
- Sentence segmentation, Sentence segmentation is segment a string of written language into its component sentences. In English and some other languages uses full stop for estimation of sentence. But in many other languages they do not use the full stop like chines. However in English this problem is not found due to the use of the full stop. Which is used for terminate the sentence. For example “*vinit went to shop in delhi.*” With the help of sentence segmentation approach we extract a sentence form handwritten data set.
- Words segmentation, it is processes in which the sentences are divide into the component words. Many algorithms are used by OCR for segment the words. The basic estimation for word segmentation is gap between the words. The space (gap) between the words can be easily found by vertically and horizontally scanning.
- Character segmentation, this is the last approach for segmentation. Character segmentation is dividing the word into character. It is tipical to divide a word into character because in cursive words create a big problem in character segmentation. On character segmentation many algorithm which is used for it with good segmentation accuracy.

There is a graphically comparison between the algorithms which is developed for line, word and character.



**FIGURE1.4 SEGMENTATION ALGORITHM GRAPH**

### 1.7.3.5 Future Extraction

The purpose of future extraction is capturing the characteristics of latter. To recognition of pattern of a latter, it is a difficult problem. The most common way of define a character by the actual image. There are three techniques which are usually used for extract such feature:

- i. Points distribution
- ii. series expansion and transformation
- iii. structural analysis

### Template matching and correlation technique

Template matching technique is different from other. It is match directly to input character image with a set of template character, which representing each possible class. The character is match with the template character class and given the best match to the pattern. This technique is simple and easy for implementation. It is used for many commercial used in ocr. But this technique has some issue when it deals with the noisy, handwritten variation and skew character.

### Point's distribution

This approach deals with the variety in handwritten style. It is extract feature based on statistical distribution of points. Some techniques are used in this area. These are:

#### Zoning

In this technique character is divide several overlapping and non-overlapping zone with rectangular circle.

#### Moments

The moments of black points about a chosen centre, for example the centre of gravity, or a chosen coordinate system, are used as features.

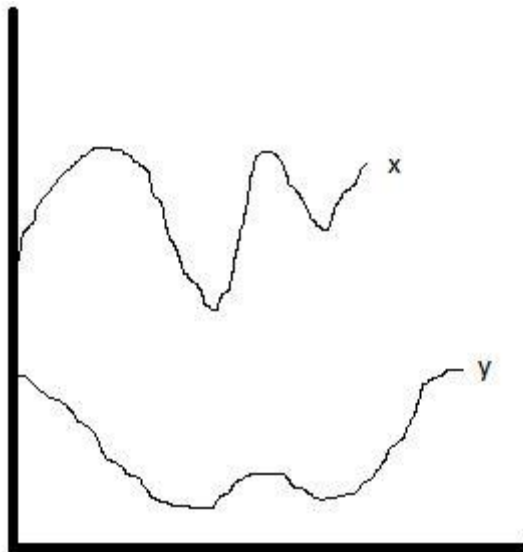
### **Crossing and distance**

In the crossing technique, the feature is extracted by number of times the character shape is crossed by the vectors with some direction. It is used for commercially used because it gives fast performance.

Distance technique is extracts the future on the basic of character distance which is passes through the vectors.

### **Transformation and series expansion**

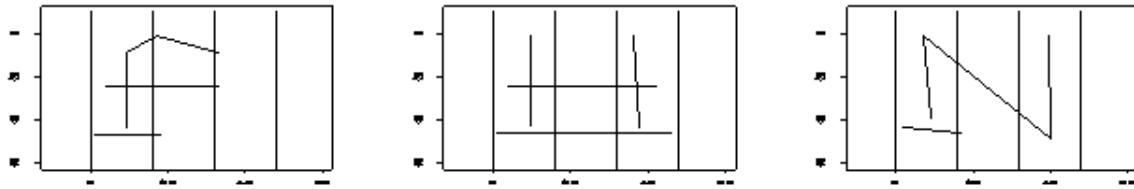
These techniques help to reduce the dimensionality of the feature vector and the extracted features can be made invariant to global deformations like translation and rotation. The transformations used may be Fourier, Walsh, Haar, Hadamard, Karhunen-Loeve, Hough, principal axis transform etc.



**FIGURE1.4 FOURIER DESCRIPTORS**

### **Structural analysis**

In structure analysis, feature is extracted according to geometric structure of a character. Structure analysis is more powerful approach for extract feature with noisy and style variation handwritten images. It extract feature with common approach like stroke, bays, endpoint and intersection between lines.



**FIGURE 1.6 STROKE EXTRACTED FROM CAPITAL LETTER N, H AND F**

## **Classification**

Classification is the method for recognize each character and map it to correct character class.

In classification, we use many methods and approach for recognition the character as classifier. Performance of classifier will be depending on its features. There are many method and rules which is used as the classifier. These methods we discuss below,

## **Decision-theoretic methods**

Under decision theoretic method come minimum distance classifier, statistical classifier and neural network. Every method will be described in briefly below.

### **Minimum distance classifier**

Minimum distance classifier works on the basis of distance. It gives good accuracy when the class are well separated. In this method used different distance measure approach but Euclidean distance is common. When the complete word is given as an input for recognition and no feature is extracted than a correlation approach is used.

### **Statistical classifier**

Statistical classifier is a probabilistic based approach. It recognizes the character on the basic of probability. According to probability method it selects the high probability character and leaves the lower probability character. Bayes classifier and HMM (hidden markow modal) are working on the basic of probability. During the recognition it assigns the character on that class which have maximum probability.

### **Neural network**

Neural network is the good approach in recently days for recognize the handwritten character. Sometime it works when the template matching is not working properly. Neural network trained by user as per demand. It recognizes the handwritten character with the help

on back propagation network. This network works with the help of feature vector, who enter in the interconnected layer. Every layer has some weight as input and transforms it as output. It used training function and adjusts weight until a desired output is not occurred.

MLP (multilayer perceptron) is the form of neural network. MLP consist multiple weighted layer of node in directed graph, in which each layer is fully connected to next one, except to input. In mlp each node is a neuron.

## **Structural Methods**

Structural method consist many method but among of them syntactic method is most useful approach.

### **Syntactic method**

In syntactic method characters are recognize by using grammatical concepts. The logic is that each class has its own grammar that defining the structure of the character. A grammar may be represented in the form of tree or strings. The structural components extracted character which is matched against the grammar of every class.

### **1.7.3.6 Post processing**

Post processing is the last step of recognition. In this process display all the recognize character in the digitalize form. It contain few process and those are,

#### **Grouping**

After the character recognition grouping play an important role when we are going to display the script. Grouping collect the recognize character and assign those character to the relevant classes. With the help of grouping character convert into words and words convert into a string.

#### **Accuracy**

Accuracy is the result of our work and it analysis the recognition success rate. Accuracy displayed in the form of percentage (%).

## CHAPTER 2

### REVIEW OF LITERATURE

---

**Rakesh and manna (2014)** has developed a recognition system for the handwritten character alphabet with the help of ANN (artificial neural network). In this approach they identify the handwritten character by observing the gradient of pixel densities. They use pixel Density Gradient (PGD) method for recognize the first five alphabetical English character.

There are many technique has been purposed previously in this field ,handwritten character recognition using row-wise segmentation technique(RST),handwritten character recognition using column-wise segmentation of image matrix(CSIM).

In the previous method are not able to achieve a good accuracy and there is a room for improvement in previous technique.

They purpose a new system by combination of both two approach row-wise segmentation and row-wise segmentation. Twelve direction methods is a feature extraction method which is depending upon the pixel of gradient. Hare pixel density gradient (PDG) extract the feature of a character which is written in different format. PGD generate a unique code for the every character. They used the first five English alphabet character , then written them on a paper .now after preprocessing binaries character 1 show the presence of image and 0 show the absence of image. They perform first row wise segmentation and after that column wise segmentation upon the input words. They used neuron and passed it multiple layer, for generate unique code for a particular character. In row wise segmentation  $R_j = \{1 \text{ if } R_{\text{tot}j} \geq \beta \text{ else } 0\}$   $R_j$  is the activation of the neuron  $j$ , produced at Neuron Layer and  $\beta$  is the threshold value which is set to 20,  $R_j$  calculates the density of dark pixels present in the row  $\beta$  is set to 20 because it is found that average pixel density is approximately equal to 20 in a particular row segment. In column wise segment the initial equation is generate  $C_j = \{1 \text{ if } C_{\text{tot}j} \geq \beta_c \text{ else } 0\}$ . Now apply pixel density gradient to generate a unique binary code and then the binary cide is decoded with the decoder and found the output. They take ten sample of each five word for the result verification and pass with the approach .They got 96% accuracy for handwritten character. We can enhance the accuracy of this method by add some neuron.

**Ankush, Sandip, Ram Sarkar et al., (2013)** has developed a handwritten word recognition system for the Bengali mixture English handwritten document. They used CMATER db 1.2.1

database for extract the English words for recognition. CMATER stand for center for microprocessor application for training education and research. CMATERdb1 is a database which contains the Bengali handwritten documents pages and in CMATERdb1.2.1 database have Bengali mixture handwritten documents pages. It's an offline recognition process.

They need to develop such system because some problem occur during the convert one language to another language, if we convert Bengali to Hindi then system become complex because it's a mix script with English.

Previous works researcher has used much other technique for recognition and segmentation for the word. In previous works heuristic and conventional method are used for segmentation and recognition. Some artificial neural network based approach also used for the recognition. Hidden markov model also used for handwritten recognition.

They perform recognition with the holistic approach, in holistic approach the complete word recognized instead of recognized character. They treat word as a single entity. They used MLP (multilayer perceptron) based classifier for word recognition; words may be cursive, touching discrete, purely discrete and mixture of cursive and discrete word. MLP consists of multiple layer of node in directed graph and each layer is fully connected to next one, except the input, each node is neuron. They contain MLP network with back propagation training algorithm for pattern recognition. Back propagation calculates the gradient of function. They extract the most frequently twelve English word occurring in the Bengali data set. The words that they are extract contain minimum frequency of 7 and maximum 68. They extract those words from the 50 pages of Bengali script. They perform segmentation upon the words up to 5<sup>th</sup> level. They contain 291 highly frequent image set for 12 words, used for learning set. For the verification of recognition they used 3 fold cross validation and achieve the average accuracy 83.24%.

Benefit of this approach is that it provides words recognition from English mixed word in Bengali document with a good accuracy.

But there is a primary limitation in this approach ,this technique not give good result when it is apply on cursive and touching discrete words.



Future work in this method is that we can improve the accuracy by adding user based dictionary with an appropriate method.

**Supachai and Buntida (2012)** have developed a model for a Bengali language recognition system for printed Thai language and English language words. OCR is applying on the large amount of data for convert it into a standard text document. In real situation documents in daily work not have a single language words, but also mix words, like Thai and English. Document may contain only a few English words in each page. They use a dictionary approach for recognize the Thai and English words. Dictionary based approach provide the batter accuracy and efficiency.

In previous work S.Tangwongsan et al develop a system for Thai-English printed documents and gives an expected result but there is some room for maintain the high level of accuracy with efficiency.

They purposed an algorithm BOCR-WP which is used for the main purpose like pre-processing, language identification, character recognition and post-processing. After the initial step of preprocessing BOCR-WP algorithm identify the language and separate words into two flow charts. One chat is use for Thai character and another one for the English character. In the initial of recognition step, they start with a word token of three or more characters. Then, an attempt is made by determining a list of predictive words from the n-gram tree based on the word token. Algorithm performs both recognition parallel such as Thai and English word recognition. They have both dictionary English and Thai, use of dictionary as to check the accuracy of words just recognized characters in the final stage. BOCR-WP performance is computed with three sets of bilingual printed documents in 35 pages or over 70,000 characters in Thai and English. The system achieves the accuracy of 99.94% and the speed of 1923 characters per second.

Benefit of this approach is, it perform faster and give the high accuracy.

**Aiquan Yuan et al., (2012)** have developed a system for handwritten word recognition. They used online segmentation technique for segmentation of words and purposed a lexicon driven approach for the recognition. They purposed two algorithms for recognition,

- a. An algorithm for recognition after segmentation.
- b. An algorithm for recognition with segmentation.

In the pre-processing they detect error around 9.62% with slant correction for better segmentation. They achieve 92% and 94.5% accuracy respectively with no slant correction and slant correction over the handwritten character recognition.

After performing recognition over the character, now they apply recognition process over the words and got 92.20% accuracy with recognition after segmentation method and 73.16% accuracy with recognition with segmentation method. The difference between accuracy shows that the wrong segmentation applied over the words. They perform all the experiments over the UNIPEN dataset.

**Neeta Nain, Subhash Panwar (2012)** has developed a system for handwritten character recognition. For character recognition they used diagonal based feature extraction technique with neural network. They used running handwritten text for recognition. They got 100% accuracy over the characters but 75% accuracy over the non-cursive handwriting and 60% accuracy over the cursive handwriting. Neural network trained over the 50 handwritten dataset which is used as an input for the recognition over the character. They trained the neural network for automatic character recognition. For performing the recognition they took a dataset of character a to z in format, upper case and lower case. With the help of neural network classifier they perform recognition on data sets. They create a diagonal based approach for all the characters. They trained the network multiple times on different input vectors.

**Mahantapas Kundu, D.K Basu et al., (2012)** has developed a system for text line extraction from Bengali and Bengali-English mixed document images. They use a database of 150 pages, 100 pages are Bengali script and 50 pages are in Bengali language mixed with English script. After collecting the dataset it is divided into two groups CMATERdb1.1.1 for Bengali script and CMATERdb1.2.1 for mixed script.

They use the neighborhood connected component algorithm upon both datasets after some preprocessing steps. They prepared ground truth images for both datasets. Connected component labeling (CCL) algorithm is used for the basic segmentation. The component is

divided into four parts type1, type2, type3 and type4. Type1 component contain small dot segments which may be or may not be connected. In type2 component have long line, type3 have long lines and type4 component have collect rest of segments. They perform algorithm on the type4 component and rest types of component concenter as the noise. Type 4 components are used to identify individual text line. In the post-processing step includes possible recall of the Type1 components, which ignored in the starting phase. Finally, there might be few cases in which some words of adjacent text lines get merged (Type #3 components). Such touching text components are carefully split into the form the actual text lines. Other type of Type #3 components considered which have maximum overlap. Result of the algorithm will be calculated by:

$$SR = (T - (U + O))/T,$$

U=number of under-segmented text lines, O=number of over-segmented text lines and T=number of actual text lines present in the document page. They achieve 90.6% accuracy in bangali database and 92.38% in the mixed dataset.

Future work is that increases the documents and adds other Indian scripts.

**Vijay patil et al., (2011)** has purposed an approach of handwritten character recognition. They used multilayer perceptron for the handwritten character recognition. They create a character matrix and network structure. Neural network used feed forward algorithm for working of neurons. They used back propagation for achieving more than 70% accuracy. They trained to the neural network for automatic character recognition. For performing the recognition they took a dataset of character a to z in format, upper case and lower case. With the help of MLP classifier they perform recognition on both data sets. They create a matrix 8x5 for all the alphabet character. They trained to the network multiple times on different input vector.

**B.gatos et al., (2010)** has purposed an approach in which some important 12 algorithm are going to compare for the best result on line and words segmentation. In 12 methods 8 are for both line and word segmentation while 4 methods are only for line segmentation. They use ICDAR2009 data set which contains English, French, German and Greek languages

documents. They extract the image and produce the ground truth image for text line and word segmentation result. It's an offline handwritten process so all the data will be offline handwritten data.

In CASIA-MST Method first split the connected component on the bases of their physical structure by Minimum Spanning Tree (MST) algorithm for text line extraction. For segmentation this method segments the words according to the gap between the connected components. It achieves the accuracy 95.68% and 84.85% for text line and word segmentation.

In CMM Method, for word detection, it calculates the average distance of the connected boxes on the bases of calculation of each line. It achieves the accuracy 98.42% and 88.91% for text line and word segmentation. In CUBS Method, in line separation is depending on run length based analysis. Word segmentation is done based on convex hull distance. It achieves the accuracy 99.53% and 86.96% for text line and word segmentation. In ETS Method, both line and word segmentation is performing by the morphological approach. In this method binarization is done by the ostus algorithm. It achieves the accuracy 86.67% and 84.93% for text line and word segmentation. In ILSP-LW09 Method, text line segmentation is done by the viterbi algorithm and for word segmentation it measures the word gap matrix with the help of SVM Classifier. It achieves the accuracy 99.05% and 94.77.85% for text line and word segmentation. Jadavpur University Method, text line and word segmentation is depend on neighborhood connected component and the connected component is found by the pixel value, where the pixel value is lower, we segment the text line and word from there. It achieves the accuracy 87.34% and 82.74% for text line and word segmentation. PAIS Method, text line are extracted on the bases of horizontal projection value of each strip and calculate the average distance between connected line. Word segmentation is done by the threshold value. It achieves the accuracy 98.52% and 90.54% for text line and word segmentation.

So the winning method is ILSP-LW09 seg for words segmentation with 94.77% accuracy and CUBS method for line segmentation with 99.53% accuracy.

**A.K, Ram Sarkar et al., (2009)** has developed a system for text line segmentation for handwritten document using neighborhood connected component analysis. The purposed a new methodology based on comparison of neighborhood connected components to know component belongs to the same text line. Components which are very small and very large to the average component height are ignored in the preprocessing step. At the time of post processing, such components are reconsidered and placed to the lines in which they belong most perfectly. The performance of purposed technique is evaluated with ICDAR2009 dataset.

Previous works uses Hough transform consider a set of points of initial image as input while the lines which is fit best to these points are calculated. A recent block based Hough Transform method takes into account gravity centers of parts of connected component. Some method used RLSA in which the value of each pixel is the sum of all pixels in the original image within a specified horizontal distance.

They set two thresholds value for remove the noise and outlier, T1 for height and T2 for width. A component has height less than T1 and width id less than T2 is ignored. They use an eight way connected component labeling algorithm for uniquely labeled and arrange a set of components for text line identification. For analysis of identified component: the average Component height (Hcavg) and width (Wcavg) is calculated as the let N be the total number of component and height take value from data set X1, X2-----Xn . Now again let the arithmetic mean of H of component to be  $\mu$ . Then,

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2},$$

A component having height from  $\mu - \sigma$  to  $\mu + \sigma$  are considered as the average height of component. Width of component will be calculated in the same manner. In this approach component having height greater than  $\delta * H_{Cavg}$  have been split into two equal parts. The post processing step based on Euclidean metric for small component which is ignored during preprocessing, fit into the suitable text line.

Future work in this approach is segmentation of text line into words.

**Marcus Liwicki and Horst Bunke (2006)** has developed a technique for the online handwritten texts capture from a white board. In whiteboard notes size and width of the character become smaller when writer move left to right during writing. They use the IAMDB dataset which contain the English handwritten data acquired from whiteboard.

They purposed the first recognition system in 2005 for whiteboard dataset but it is based on offline handwritten recognition. The online data transferred into form of image of text line. Some information may be loses during the transformation process. So in 2006 they purposed a system for online whiteboard data.

They used hidden markov models (HMM) for the classification and N-gram model used for preprocessing. Online recognition is representing the moment of pen tip that could be electronic pen or white board pen. This approach is interesting because they recognized words which is write on the white board and challenges is that people is stand, rather than sit, during writing the arm does not rest on a table, handwriting on a whiteboard is different from handwriting produced with a pen on a writing on the paper. On the white board size and width of the characters become smaller the more the writer moves to the right. In the preprocessing normalized the words because the style of writer is different with respect to skew, slant, height and width of the character. Data are collected from the IAMonDB , it is a large online handwritten database consisting of handwritten whiteboard notes. They used 58 characters from the character set in which include all small and capital latter together. They used Baum-Welch algorithm for the training set and in the recognition viterbe algorithm is used for finding the most probable word sequence. They also used the N-gram language model which is based on the observation, we able to guess when we are reading the text; we can say probability of next word is depending on the previous one. They achieved word recognition rates of 67.3% on the test set when no language model is used, and 70.8% by including a language model.

**Alessandro L. Koerich, Alcué S.Britto (2006)** has developed a system for handwritten words in form on combination of HMM words classifier and segment neural network (SNN) classifier. They purpose a novel approach with low level and high level feature for words recognition. They recognized the word and character with the help on combination of method. For batter recognition result they offer two categories of feature: high level feature

have global, histogram and segmentation feature which comes from word segmentation and generate variable length feature. Low level feature have feature which comes from character.

They use HMM are for the high level feature extraction and NN used for the low level feature extraction. In high level feature a word image of high level features are extracted from loosely segmented words. Such features are used with an HMM word classifier in a lexicon-driven approach. A lexicon driven approach is the best process of handwritten recognition suitable for real time application.HMM gives the best output with list of n-best recognition hypothesis. They consider feature at the grapheme level because of arrange the clustering letter into classes.HMM uses the viterbi algorithm for recognize the character from words. SNN used a standard MLP character classifier. This classifier performs the character recognition task. Than character arrange into a unique class like, A and 'a'. Now the network produced 26 outputs, one for each character. Then combine both two approach and produced output. High level feature achieve the 32% words error rate and low level feature achieve the error rate 22%. So the aggregate accuracy is 71%.

**Victor Lavrenko et al., (2004)** has develop a system in which complete word is recognize rather than segmentation words into smaller pieces and then perform recognition separately. In holistic approach complete word recognize at once. This approach is parallel for the human reading. They used the historical data of a single author. It's a offline handwritten recognition approach. They said if we apply segmentation on word into character than we got a poor result.

In the previous works they showed a recognition rate of about 60% for vocabulary sizes ranging from 2703 to 7719 words. They also maintain a discussion of researchers - these varied from a recognition rate of 42.5% recognition rates obtained by other for a 525 word vocabulary and 37% for a 966 word vocabulary reported in to a recognition rate of 55.6% in a1600 word vocabulary.

They used hidden markow model for the recognition. HMM model is used for find the hidden state of the word and estimated word bigram frequency reported by. They purposed HMM for recognition span and is able to keep in mind only the last word she, they assume that the author has an extremely short memory has written. Given that the last word was

$w_{j-1} \in V$ , the author picks the next word  $w_j$  according to some probability distribution  $P(w_j | w_{j-1})$ . In Hidden Markov Model the words  $w_1, \dots, w_n$  represent the state sequence. Each state depends only on the previous state  $P(w_j | w_{j-1})$ . They take a set of experiment for the test of effectiveness of our model, so they chose a writer gorge manuscript. They extract the 20 page from the latter written by Gorge. In this process the word fort and Fort are treated two different word, now apply 20 fold cross validation on the 20 pages. They use the word error rate for that word those are not recovered exactly as they are written in the manual script. They achieved 65% accuracy with this system.

Accuracy of this system will be improve by improve the vocabulary terms and improve the quality of n-gram.



#### 3.1 PROBLEM FORMULATION

Our approach is design for multiple handwriting style. The main purpose of our system is used based approach for recognition the words. . It's an offline handwritten words recognition method. Offline handwritten recognition is an approach that is apply on the old data it means perform recognition after written data. In present work we perform recognition on the English words which is collected from multiple people in different handwriting style. We use analytical approach for recognition to the words. Analytical approach recognize to the word character by character. We uses neural network with back propagation approach for recognition word. Because we working with an analytical approach so neural network recognize the word character by character. If some character is not recognized perfectly than we uses user based interface. In user based interface, the user gives an input by the keyboard to character which is not recognized by neural network.

#### 3.2 OBJECTIVES

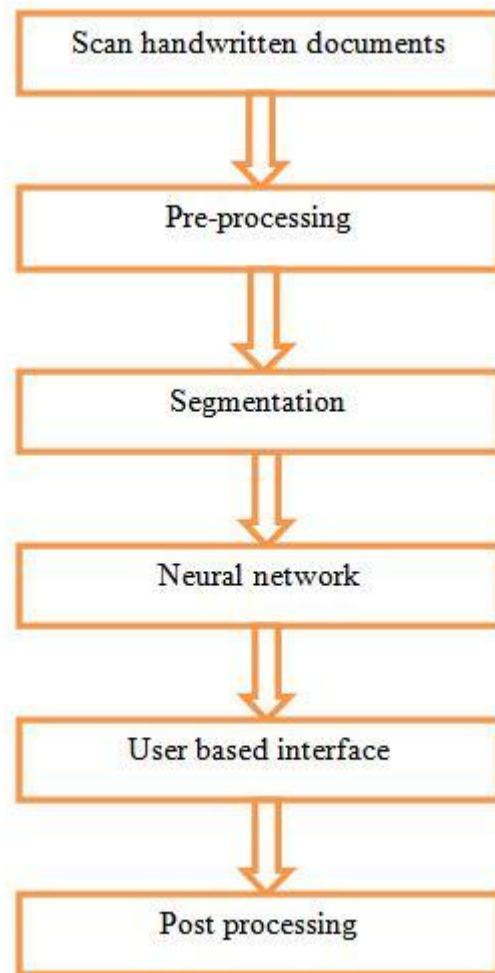
Objective of our work is to,

- I. Recognize maximum words from dataset.
- II. Improve accuracy of optical character recognition.
- III. Enhance the performance of optical character recognition.
- IV. Use our approach on multiple people handwritten document.

#### 3.3 RESEARCH METHODOLOGY

In our methodology we represent a method for recognize English handwritten words from different people handwritten dataset. In the first step in recognition is scanning the documents by high quality flatbed scanner with 300dpi gray scale image resolution. Now all

the scanned pages are stored in the bitmap file format. Bitmap file format is able to stored 2D images with arbitrary width, height and resolution and also in different color.



**FIGURE 3.1 FLOW CHART OF PROPOSED WORK**

### **Preprocessing**

In processing first of all the scanned document are binaries. For binarization of an image we use the technique those are developed for image binirazition. In binarized image pixels are show as '0' and '1' format.

Now we remove the noise from binary image. Noise may be paper noise, outlier and small dots. We apply median filter noise removal technique for remove slant and paper noise. Now we normalize the image it means we normalize skew and slant from the handwritten document with relevant techniques.

## **Segmentation Process**

In segmentation, the initial work is to segment the line from document page. Here we use an approach for text line segmentation from handwritten document image with the help of connected component analysis. We segment handwritten text line on the bases of pixel value of connected component. If two lines are connected with each other then we calculate the pixel value of both line and in between two lines where the value is less than average pixel value than segment line from there.

Now the second phase of segmentation is word segmentation, in this phase we use connected component segmentation algorithm. Basically word segmentation is depends upon the gap between of two words. Gap metric between two running connected component is define by the SVM classifier. Threshold value is given by calculating the all gap metric values from the documents

## **Recognition Process**

In this process we purposed an user based intelligent approach for recognizing handwritten word. We will explain it step wise step,

**Step 1:** In the initial step of recognition we used neural network for recognition of character with the help of template matching. Neural network try to recognize all cursive and non-cursive words in the form of analytical approach with back propagation algorithm. But due to some words structure problem some character is not recognize from the words.

**Step 2:** Now after performing neural network on user based data, the neural network display those character which is not recognize.

**Step 3:** Now we create an user based interface, our user based interface is work on the unrecognized character because our approach is user based method. So the user gives the

input from the key board to those character which is not recognized and display over the screen.

We trained our neural network, so when the user gives an input character from key board than the neural network mapping with recognize text file to the initial input image line by line and understand which character is missing from the word and put the key board character where it is missing.

**Step 4:** Now we got recognize and complete words with the help of user based input interface. We increase the accuracy 8 to 9% with the help of this user based approach.

## CHAPTER 4

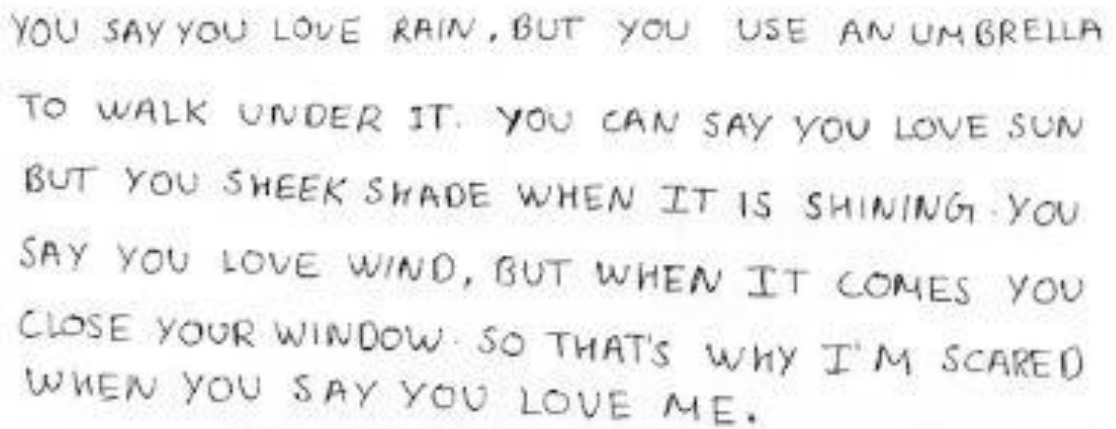
### RESULT AND DISCUSSIONS

---

**Testing and experiment results:** We collect the data from different people with different age gape. The data set which is we collect in different form,

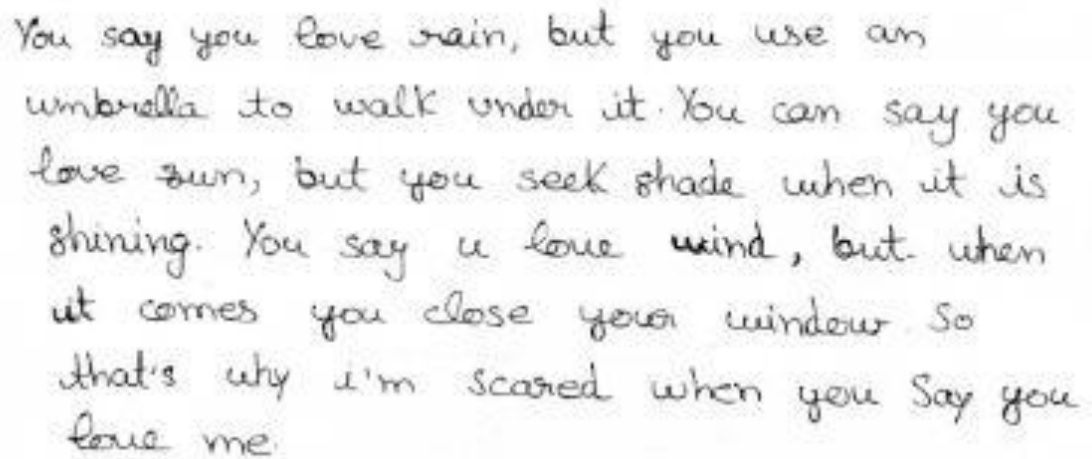
- i. Collect the handwritten data in upper case letter.
- ii. Data set in the lower case letter form.
- iii. Data set in mixed form it means upper case and lower case.
- iv. Collect A to Z and a to z character from different people with age difference.

The data is,



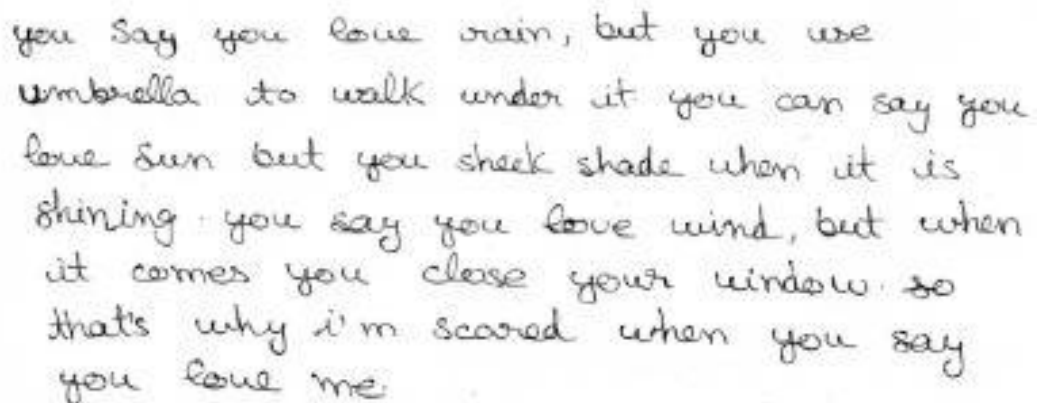
YOU SAY YOU LOVE RAIN , BUT YOU USE AN UMBRELLA  
TO WALK UNDER IT. YOU CAN SAY YOU LOVE SUN  
BUT YOU SHEEK SHADE WHEN IT IS SHINING. YOU  
SAY YOU LOVE WIND, BUT WHEN IT COMES YOU  
CLOSE YOUR WINDOW. SO THAT'S WHY I'M SCARED  
WHEN YOU SAY YOU LOVE ME.

**FIGURE5.1 HANDWRITTEN SET UPPER CASE**

A photograph of a piece of paper with handwritten text in a cursive script. The text is written in a mix of uppercase and lowercase letters. The content is a short paragraph about the irony of loving weather while avoiding it.

You say you love rain, but you use an umbrella to walk under it. You can say you love sun, but you seek shade when it is shining. You say u love wind, but when it comes you close your window. So that's why i'm scared when you say you love me.

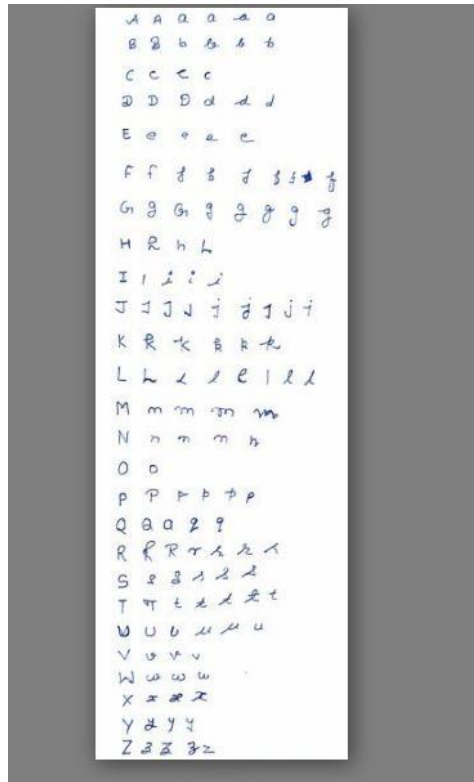
**FIGURE5.2 HANDWRITTEN SET IN UPPERAND LOWER CASE**

A photograph of a piece of paper with handwritten text in a cursive script. The text is written entirely in lowercase letters. The content is the same short paragraph as in Figure 5.2.

you say you love rain, but you use umbrella to walk under it you can say you love sun but you seek shade when it is shining you say you love wind, but when it comes you close your window so that's why i'm scared when you say you love me

**FIGURE5.3 HANDWRITTEN SET IN LOWER CASE**

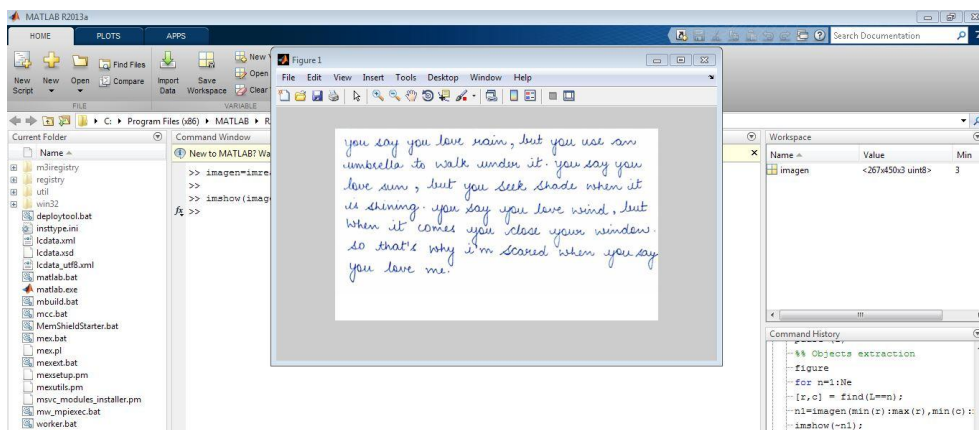
We trained to the neural network over different people handwriting. We collect different user capital letter handwritten A to Z and a to z in small letter and trained network over it. Now we run neural network over the multiple handwritten data and got an average accuracy 83 % without user based interface. The accuracy will be increase 12% to 15% with the user based interface.



**FIGURE 5.4 NEURAL NETWORK TRAINING DATA SET**

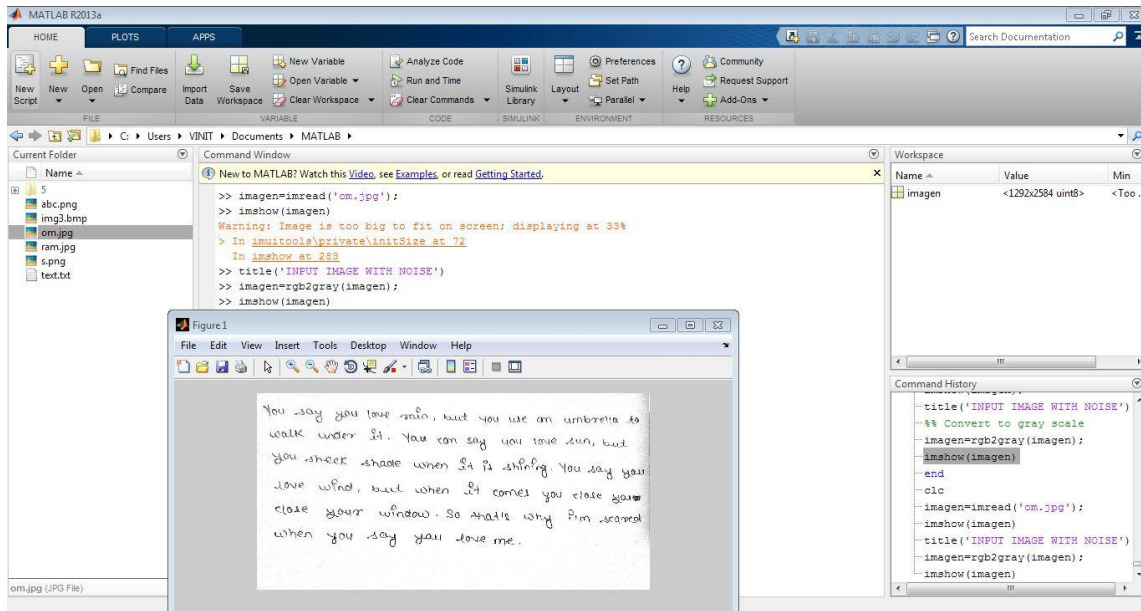
Now we are going for recognition of those data sets,

When we are performing recognition the initial step is provide an input image with the help of matlab code.



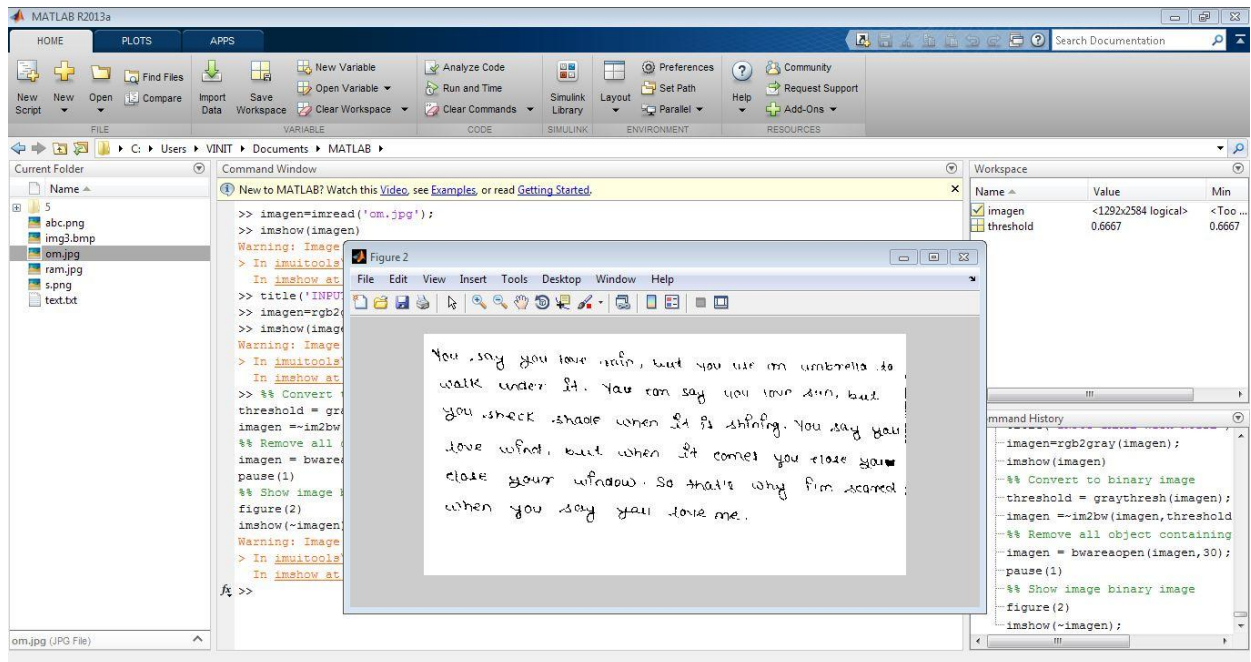
**FIGURE 5.5 INPUT HANDWRITTEN TEXT IMAGE**

After input image we binarized it by `image =rgb2gray(a);`



**FIGURE 5.6 BINARIZED IMAGE WITH NOISE**

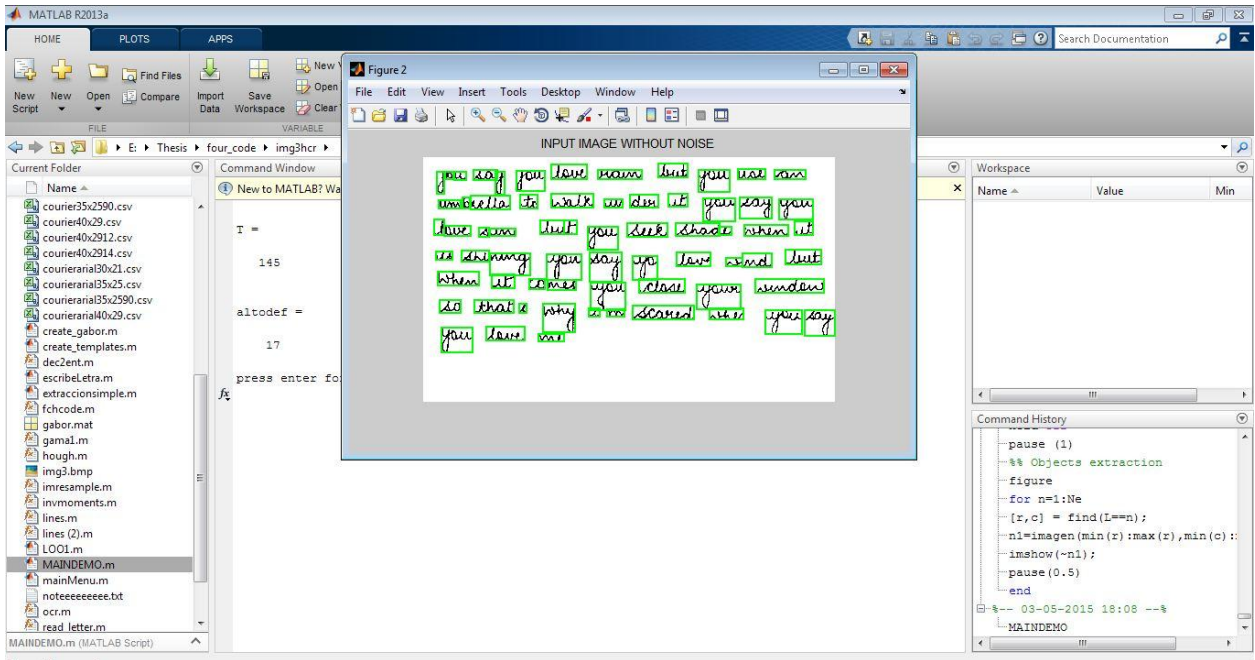
But binarized image have some noise, than we have to remove it.



**FIGURE 5.6 REMOVE NOISE FROM IMAGE**

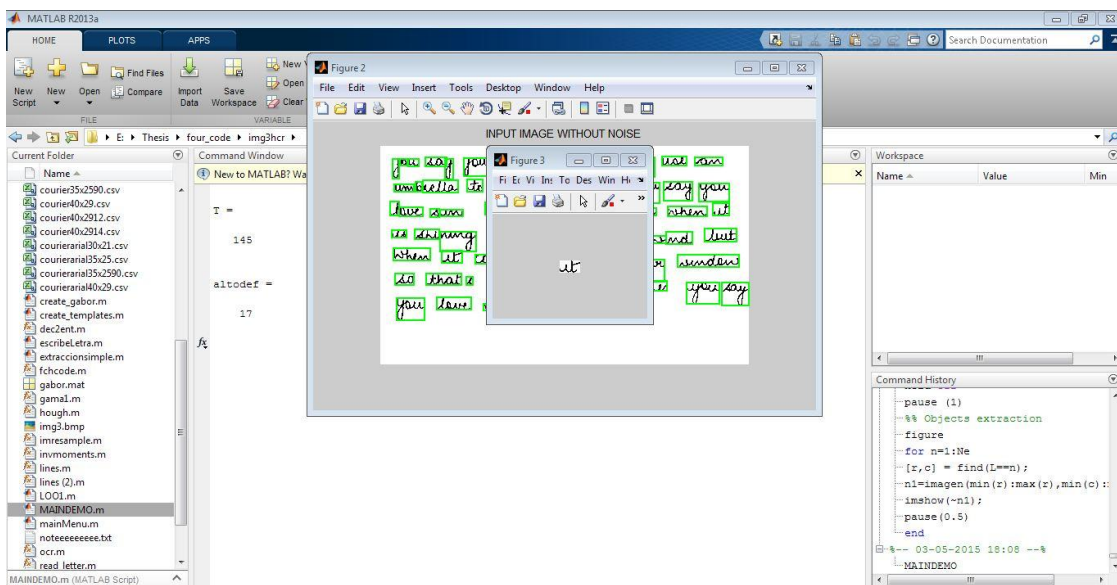


Now after remove noise the next step is segmentation to the image text,



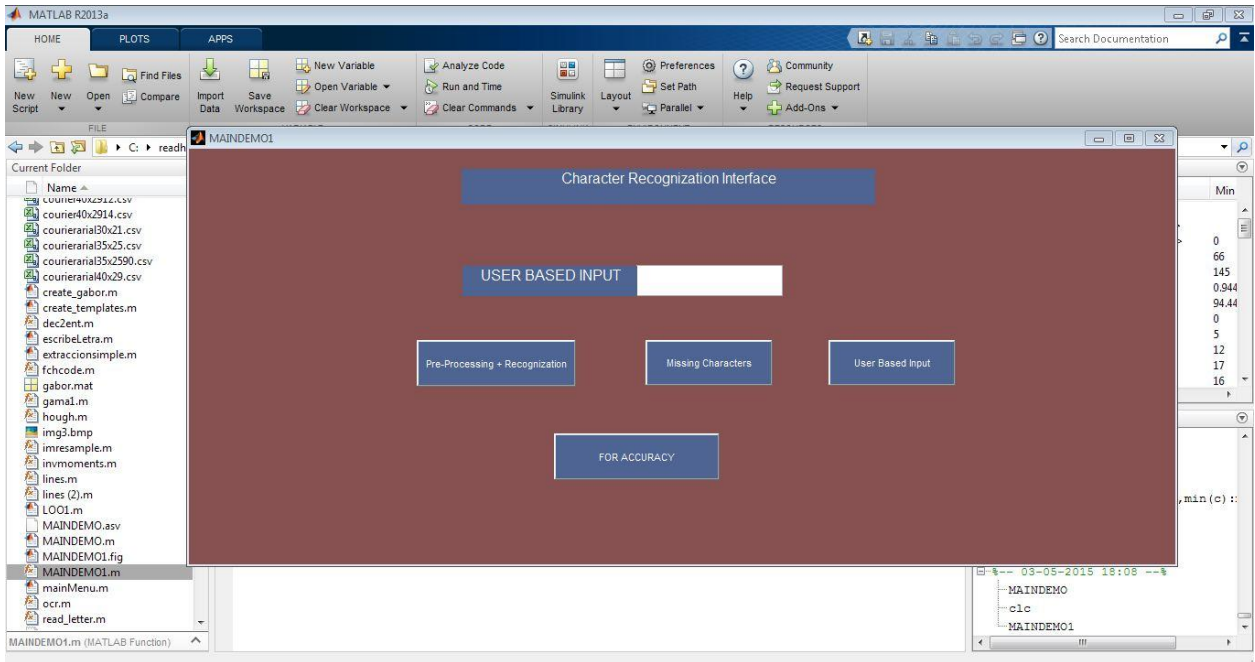
**FIGURE 5.7 TEXT SEGMENTATION WITH MATLAB CODE**

We done the segmentation, now we will extract the words or character from segmented text image for recognition,

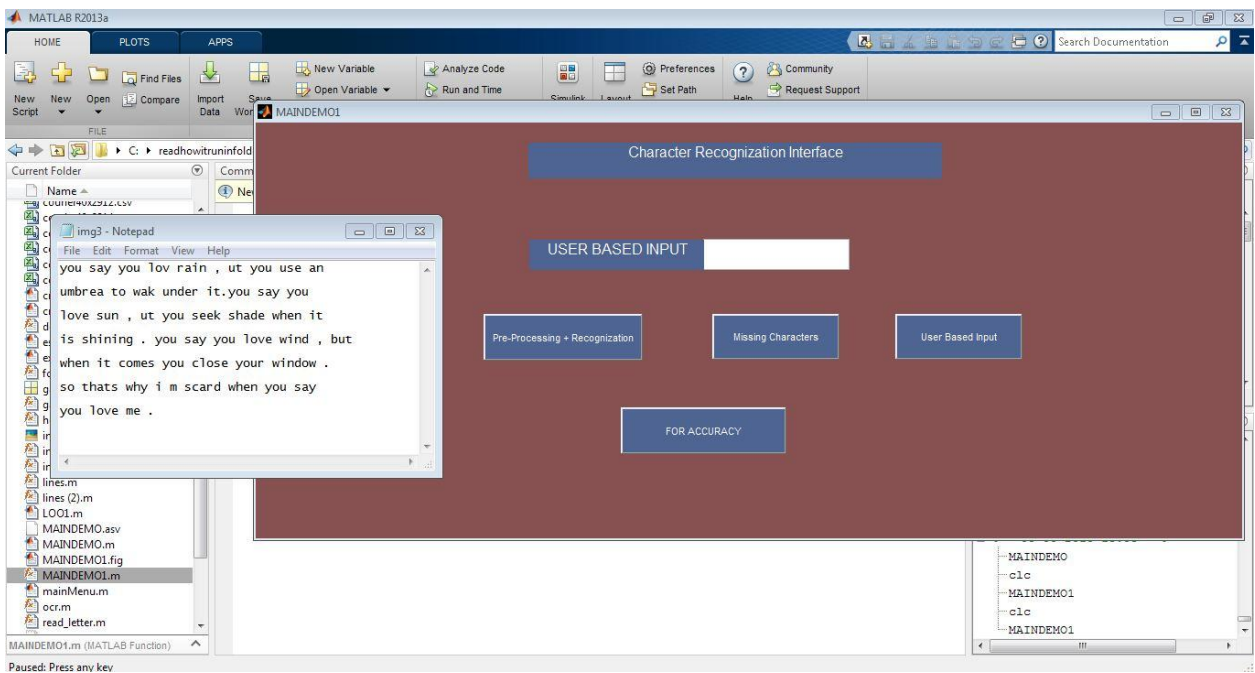


**FIGURE 5.8 EXTRACT WORDS AND CHARACTER FROM SEGMENTED IMAGE**

In this step the recognition is start with analytical approach by neural network. We create an interface for recognition.

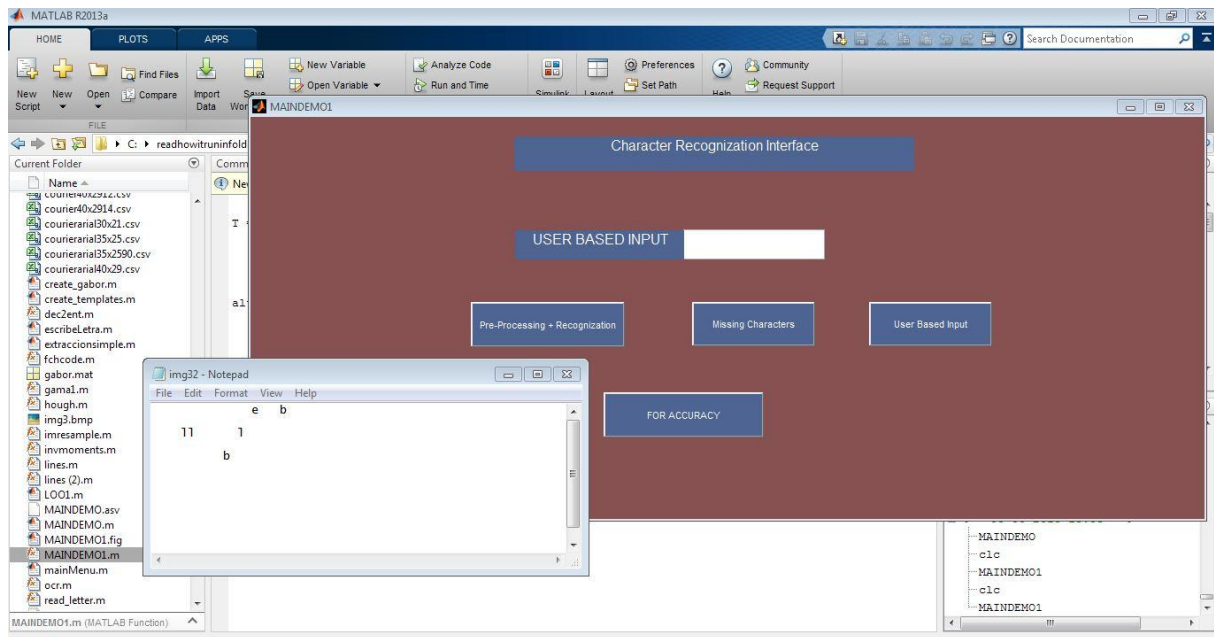


**FIGURE 5.9 RECOGNITION INTERFACE**



**FIGURE 5.10 RECOGNITION WITH NEURAL NETWORK**

After recognition our system display those character which is not recognize,

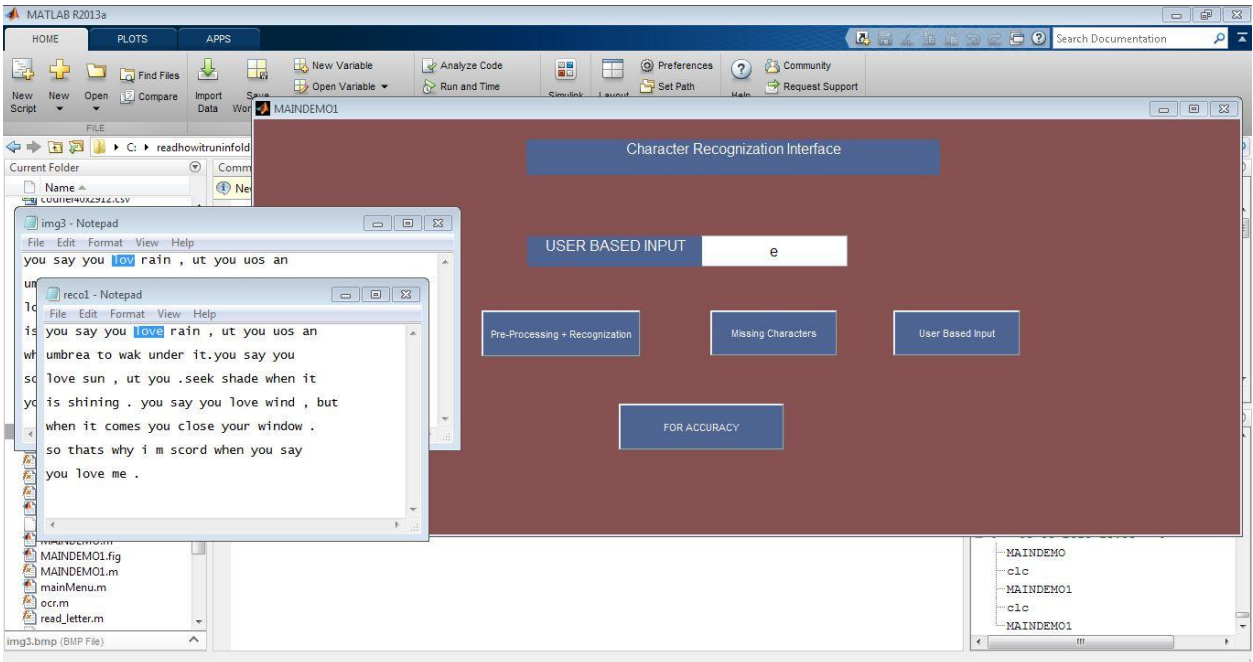


**FIGURE 5.11 DISPLAY CHARACTER WHICH IS NOT RECOGNIZED**

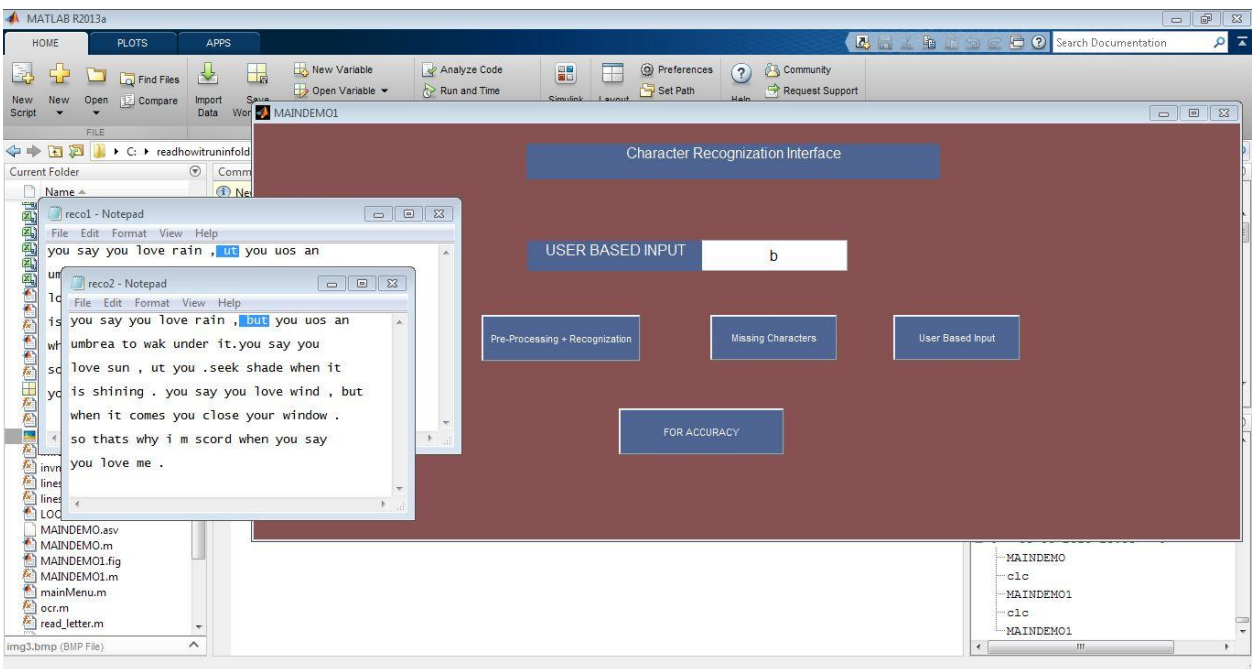
Now we created a user based interface, by using a user based interface we give an input to the system with the help of a keyboard to those characters which are not recognized. After providing the character input by keyboard, the user-based input fills the missing character into the word automatically.

For example, in the below snapshot we show that the character 'e' is missing from the word 'love' and the word displays as 'lov'. When the user provides an input character 'e' from the keyboard which is missing, then the missing character automatically fills the word.

Repeat the same process until all missing characters are not filling the words.

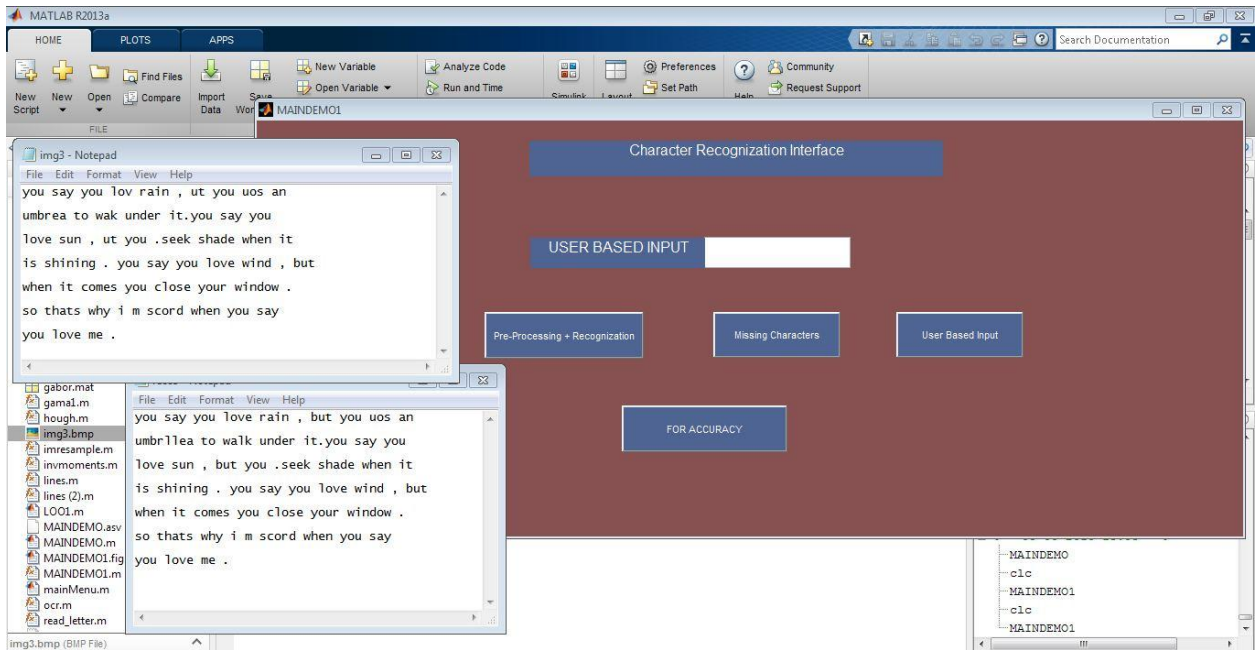


**FIGURE5.12 USER BASED INPUT CHARACTER ‘e’ WHICH IS NOT RECOGNIZED AND DISPLAY THE RESULT**



**FIGURE5.13 USER BASED INPUT CHARACTER ‘b’ WHICH IS NOT RECOGNIZED AND DISPLAY THE RESULT**

After completed the filling process the system displayed the final text file.



**FIGURE 5.14 COMPARIN BOTH FILE**

### FUTURE SCOPE AND CONCLUSION

---

#### FUTURE SCOPE

Optical Character Recognition (OCR) is an open field for the handwritten recognition. It's a challenging task to recognize the handwritten words with a good accuracy. In this work we perform recognition on English words which is collect from different people handwriting dataset. Over the dataset we apply present approach and recognize the character with analytical approach. This approach is helpful for that researcher who wants to convert a different people handwritten data into digitalize form. So this work is to improve the performance of handwritten recognition. It is very impressive method for the handwritten.

#### CONCLUSION

In the present work we represent an analytical approach and recognize the English handwritten words from different user's dataset. We use neural network with back propagation approach for recognition of handwritten character which is written in different-different style. A neural network based approach is used for recognition words. Neural network recognize the character from the words but some character is not recognize by neural network. For unrecognized character we a user based interface. By interface we provide an input from keyboard to those words where the character is missed. So with the help of key board we filled missed character. We can also say that it is a hybrid approach with multiple approaches. Mostly we see that to recognize cursive, mixture of cursive, discrete and touching discrete is an open challenge in the field of handwritten recognition. So we will recognize all these types of words by the present methodology and get a desired output.

## CHAPTER 6 REFERENCES

---

### I. Books

James Allen (2005)“Natural Language Understanding” Published by Pearson Education Singapore.

A.Bharti, Vineet Chaitanya, Rajeev Singal (1995) “Natural Language Processing: A Paninian Perspective” PHI Learning Private Limited, New Delhi.

### II. Research Paper

RK Mandal, NR Manna (2014) “handwritten English character recognition using pixel density gradient method”, International Journal of Computer Sciences - ijseonline.org

A. Acharyya, S. Rakshit, R. Sarkar, S. Basu, M. Nasipuri (2013) “Handwritten word recognition using MLP based classifier: a holistic approach” International Journal of Computer Sciences Issues, Vol., Issue 2.

S. Tangwongsan, B. Suvacharakulon (2012) “OCR with Word Prediction Technique for Bilingual Documents”, IEEE/ACIS 11<sup>th</sup> International Conference on Computer and Information Science.

Aiquan Yuan, Gang Bai, Po Yang, Yanni Guo, Xinting Zhao (2012) “Handwritten EnglishWord Recognition based on Convolutional Neural Networks” IEEE DOI 10.1109/ICFHR.2012.210

Neeta Nain Subhash Panwar(2012)” Handwritten Text Recognition System Based onNeural Network” academepublish.org –Journal of computer and information technology vol.2, No.2

Vijay Patil and Sanjay Shimpi (2011)“Handwritten English character recognition using neural network”, Elixir Comp. Sci. & Engg. PP. 5587-5591.

R. Sarkar, N. Das, S. Basu, M.Kundu, M.Nasipuri, D.K Basu (2011) “CAMTERdb1: a database of unconstrained handwritten Bangla and Bangla-English mixed script document image” @Springer.

B. Gatos, N.Stamatopoulos, G.Louloudies (2010) “ICDAR2009 handwritten segmentation contest”@Springer.

A.k, P.Choudhury, R. Sarkar, N. Das, S. Basu, M.Kundu, M.Nasipuri, N.Das (2009) “Text Line Segmentation for Unconstrained Handwritten Document Images Using Neighborhood Connected Component Analysis”@Springer, pp.369-374.

M. Liwicki and H. Bunke (2006) “HMM-based on-line recognition of handwritten whiteboard notes,” in Tenth International Workshop on Frontiers in Handwriting Recognition.

A. L. Koerich, A. S. Britto Jr, L. E. S. de Oliveira, R. Sabourin (2006) “Fusing High- and Low-Level Features for Handwritten Word Recognition” in Workshop on Frontiers in Handwritten Recognition.

Victor Lavrenko, Toni M. Rath and R. Manmatha (2004)“Holistic Word Recognition for Handwritten Historical Documents” in First International Workshop on Document Image Analysis for Libraries, IEEE.



## **CHAPTER 8**

### **APPENDIX**

---

OCR: - Optical Character Recognition

NLP: - Natural Language Processing

HWR: - Handwritten Recognition

