LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

An approach to misclassification detection and correction in Optical Character Recognition

A Dissertation Report

Submitted

By

Aman Kumar

(11304832)

To

Department of Computer Science & Engineering

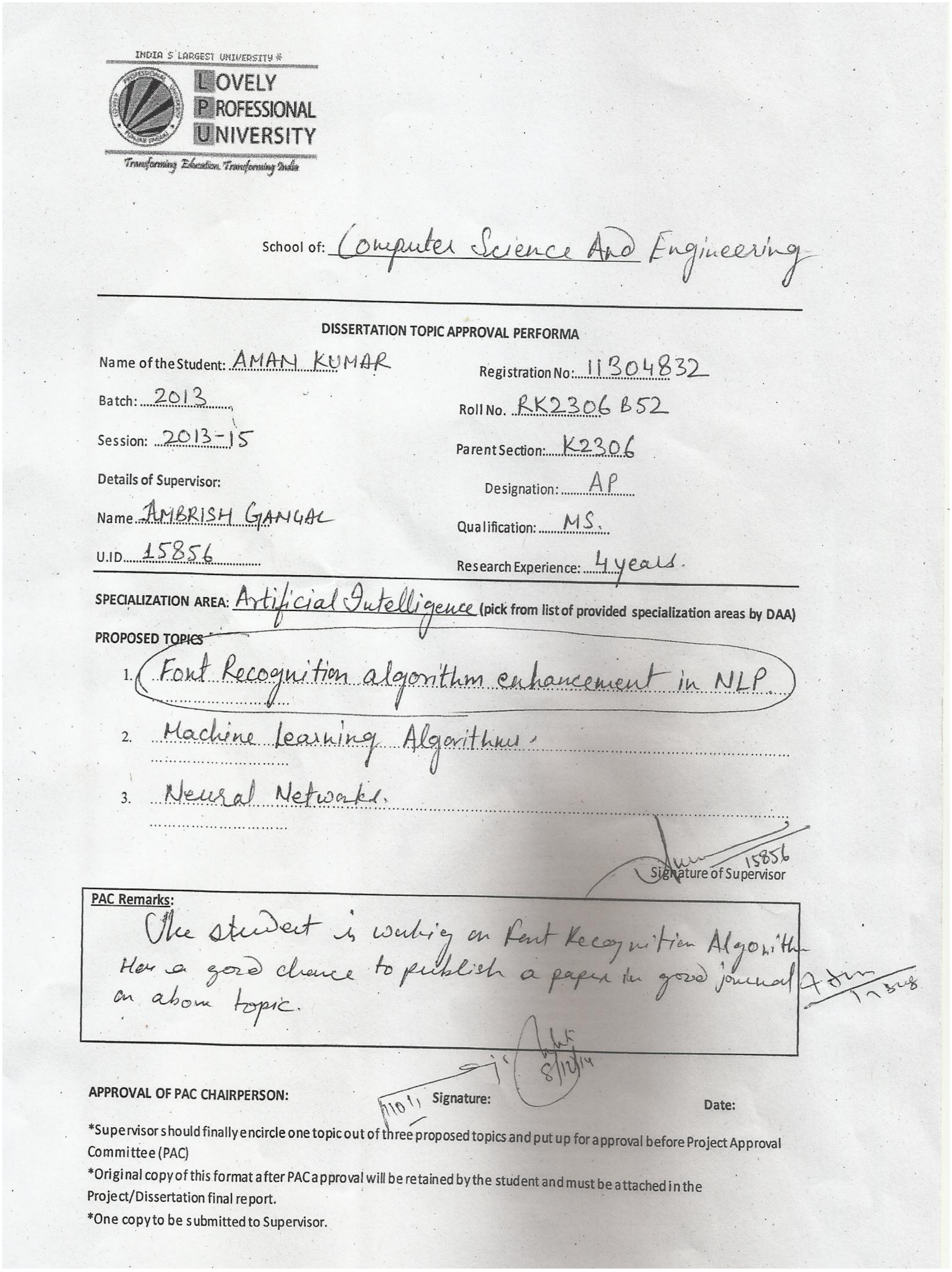
In partial fulfillment of the requirement for the

Award of the degree of

Master of Technology in Computer Science

Under the guidance of
Ambrish Gangal (15856)

(May 2015)



INDIA'S LARGEST UNIVERSITY*

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: Computer Science And Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: AMAN KUMAR Registration No: 11304832

Batch: 2013 Roll No. RK2306B52

Session: 2013-15 Parent Section: K2306

Details of Supervisor: Designation: AP

Name AMBRISH GANGAL Qualification: MS.

U.ID 15856 Research Experience: 4 years.

SPECIALIZATION AREA: Artificial Intelligence (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Font Recognition algorithm enhancement in NLP
2. Machine learning Algorithms.
3. Neural Networks.

15856

Signature of Supervisor

PAC Remarks:

The student is working on Font Recognition Algorithm Has a good chance to publish a paper in good journal on above topic.

APPROVAL OF PAC CHAIRPERSON: 11011 Signature: 8/12/14 Date:

*Supervisor should finally encircle one topic out of three proposed topics and put up for 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

This paper introduces a complete Optical Character Recognition (OCR) framework for picture/illustrations inserted literary records for handheld gadgets.. At first, we have showed the misclassification encountered in previous works. Then we have discussed our implementation where content are separated and skew rectified. At that point, these locales are binarized and segmented into lines and characters. Characters are gone into the acknowledgment module. Trying different things with an arrangement of pictures, we have accomplished a greatest acknowledgment precision.

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant guidance crowned our efforts with success.

I sincerely express our deep gratitude to the management of our college for giving us liberty to choose and to work on the most relevant project i.e. "**An approach to misclassification detection and correction in Optical Character Recognition**". I am thankful to **Dr. Dalwinder Singh** (HOD, CSE DEPT.) for ensuring that we have a smooth environment in the university by providing us with the best suitable mentors according to our field. I would also like to thank the Research and Development department (R&D department) for providing the opportunity to conduct the research work.

I would like to thank my guide **Ambrish Gangal**, **Assistant Professor, CSE Department,** who encouraged and insisted us in the formulation of problem definition & without his valuable guidance and constant inspiration it would have been difficult for us to prepare this project report.

DECLARATION

I hereby declare that the dissertation proposal entitled, "**An approach to misclassification detection and correction in Optical Character Recognition**" submitted 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: __________

Aman Kumar

Reg. No. 11304832

CERTIFICATE

This is to certify that **Aman Kumar** has completed M.Tech dissertation report titled "**An approach to misclassification detection and correction in Optical Character Recognition**" under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation report has ever been submitted for any other degree or diploma.

The dissertation report is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science and Engineering.

Date:

Signature of Guide
Name: Mr. Ambrish Gangal
UID: 15856

Table of Content

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The Natural Language Processing is an area of Computer Science to provide better communication into computers and Human Natural Language. Human Level Natural Language is Artificial Intelligence -main problem. The Natural Language Processing solving central AI problem to making computers as intelligent as people. The Natural Language Processing is development of AI that upgrades computers to give the knowledge and it understands online data and applies what the learned in real world. The Natural Language Processing makes computers more capable receiving and giving more and more instructions. The main goal of Natural Language Processing evolution is to measure more and more qualities of a system.

The AI is a forcing function of Computer Sciences. It pushes the boundaries of Computer Science by demanding more and more from machine. A machine can give the result like a human being. So, Natural Language processing is a term of AI that related to Machine Learning and Knowledge representation. Natural Language Processing is important topic of Internet. There is lots of information on the Web in the form of text. It is important concern of to update the information of this task. This is the motivation for understanding NLP, its tools, technique and principles. So, Natural Language Processing related to the area of human-computer interaction. The Natural Language Processing complete major task such as automatic summarization, machine translation, morphological segmentation, text generation, questioning Answering, topic speech understanding and segmentation. The main objectives of Natural Language Processing are to provide conclusion by the system to gain the knowledge of computer system and it will able to take the decision like a human being in more accurate way.

Theoretically regular dialect transforming is "a hypothetically propelled accumulation of computation strategies for determining and speaking to regularly happening writings at one or more level of semantic investigation with the end goal of accomplishing to process the human-like dialect for a scope of archives or applications." The procedure of PC examination of data gave in a human dialect (characteristic dialect), and change of this information into a helpful type of representation. The field of NLP is fundamentally concerned with getting PCs to perform helpful and intriguing undertakings with human dialects. The field of NLP is optionally concerned with helping us go to a superior comprehension of human dialect.

The objective of NLP is to outline and fabricate programming that will investigate, comprehend and produce dialects that people utilize characteristically. Common dialect understanding frameworks proselyte tests of human dialect into more formal representations, for example, parse trees or first request rationale that are simpler for PC projects to control. In principle, common dialect transforming is an extremely alluring technique for human – PC association. Early frameworks working in confined "pieces universes" with limited vocabularies, worked amazingly well, driving scientists to unnecessary confidence, which was soon lost when the frameworks were stretched out to more practical circumstances with true vagueness and unpredictability. On the off chance that one goes to history of NLP, then 4 one would discover a paper titled "Processing Machinery and Intelligence" and this was composed by Alan Turing distributed in 1950 as Turing [1950]. These days, the substance of this paper are known as the "Turing test" as a basis of knowledge. This measure relies on upon the capacity of a PC project to mimic a human in a constant composed discussion with a human judge. In 1954, sixty Russian sentences were completely deciphered into English and that investigation was known as the "Georgetown test". There was a moderate advance in this field till 1980, when the first measurable machine interpretation framework was produced. In late 1980, machine learning calculations for dialect handling were presented. This changed the idea of NLP. This expanded the computational force. IBM examination gatherings were working in the field of machine interpretation and these kind of frameworks had the capacity interpret a few reports effectively. These days, the anxiety is on less administered or unsupervised learning calculations. Anyway, this kind of unsupervised learning is substantially harder to attain. NLP has critical cover with the field of computational etymology, and is regularly viewed as a sub field of computerized reasoning. There are numerous utilizations of Natural Language handling grew throughout the years. The primary are content based applications, which includes applications, for example, hunting down a certain point or a pivotal word in a vast report, making an interpretation of one dialect to another or abridging content for diverse purposes. In this way, common dialect transforming is the branch of software engineering which arrangements to change over reports written in one of the characteristic dialects to machine justifiable organization. For this kind of transformation, one or mixes of coding languages are utilized. Here inquiry emerges that whether codes vary than common dialect. Yes, these vary. The term regular dialect is utilized to recognize human dialects, (for example, English, Spanish, and so on.) from formal or scripting languages, Here is some intriguing correlation between regular dialect and coding languages. As common dialect alluded to as human dialect whereas, code is a dialect adequate to a PC framework. In common dialect, every word has a positive importance and can be turned upward in a lexicon. In the comparative way, every single programming language has a vocabulary they could call their own. Every expression of that vocabulary has a clear unambiguous significance, which can be turned upward in the manual implied for that dialect. The fundamental distinction between a characteristic dialect and code is that regular dialect has expansive vocabulary however most scripts utilize an exceptionally restricted or limited vocabulary. This is on account of a programming dialect, by its temperament and reason, does not have to say excessively. Each issue to be explained by PC must be separated into discrete (straightforward and separate), sensible steps, which essentially involve four crucial operations like information and yield operations, development of data inside the CPU and memory, and legitimate or examination operations. Every common dialect has an orderly strategy for utilizing the words and images of that dialect, which is characterized by the sentence structure principles of the dialect. Also the words and images of a programming language should likewise be utilized according to set guidelines, which are known as the linguistic structure tenets of the dialect. Individuals can utilize poor and erroneous vocabulary and language structure, and still make them caught on. However on account of programming language, one must stick to the precise linguistic structure guidelines of the dialect, if one needs to be seen effectively by the PC. Yet, no PC is equipped for amending and reasoning significance from inaccurate direction. In this way, scripts are littler and less complex than common dialects, yet they must be utilized with extraordinary exactness, at exactly that point right results can be normal.

1.1. Font Recognition

Font Recognition is a concept of identify and investigation of documents. Today's the document images such as scanned text documents are widely used. For example, there are many documents that are scanned and reserved electronically in some libraries and computers are not able to search or understand context of such document. So, font recognition is identifying this type of text image in one script to another script.

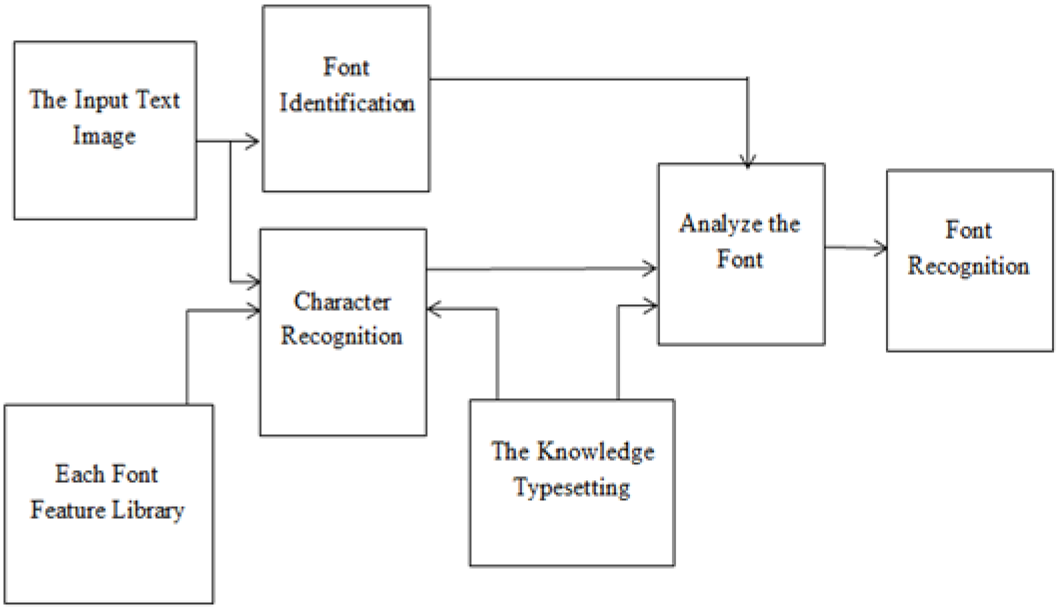


Figure 1.1 Font Recognition

In the figure 1.1 describe how font recognition is described.

- The input scanned image of any text that is used to recognized.
- For single Font recognition take the feature dictionary of font, it is important for identify the correlated fonts.
- The term Font Identification gives the guidance recognizes the first character of input text image or text block.
- The matching result is based on the guidance of font then takes to analyze the font.
- If the character is unknown then, a new direction font is selected with the help of knowledge typesetting.
- Analyze the Font; the font will be recognition of particular text block.

1.2 Character Recognition

Character recognition is the term, which covers a wide range of machine acknowledgment of characters in different application spaces. The concentrated exploration exertion on the field of character acknowledgment was not just on account of its test on recreation of human perusing, additionally, on the grounds that it gives productive applications, for example, the programmed handling of mass measure of papers, moving information into machines and web interface to paper archives. A character acknowledgment framework can be either "online" or "disconnected from the net. As indicated by the method of information obtaining, character acknowledgment philosophies are arranged into two frameworks as Online Character Recognition Systems and Offline Character Recognition Systems.

1.2.1 Online Character Recognition

Online character recognition is the process of perceiving penmanship, recorded with digitizer, as a period arrangement of pen directions. It catches the fleeting and element data of the pen direction. Uses of on line character acknowledgment frameworks incorporate little handheld gadgets, which require a pen just. PC interfaces complex sight and sound frameworks, which utilize various data modalities including filtered reports, console and electronic pen. These frameworks are valuable in social situations where discourse does not give enough protection. Pen based PCs, instructive programming for showing penmanship and mark verifiers are the samples of famous apparatuses using the on line character acknowledgment strategies.

1.2.2 Offline Character Recognition

Offline character recognition is the process of changing over the picture of composed record into bit design by an optically digitizing gadget, for example, cam or optical scanner. The acknowledgment is done on this bit design information for machine printed or written by hand content. Utilizations of disconnected from the net acknowledgment are expansive scale information handling, for example, postal location perusing, check sorting, and office mechanization for content section, programmed review and distinguishing proof. Logged off character acknowledgment is an imperative instrument for production of the electronic libraries. Additionally, the across the board utilization of web requires the use of disconnected from the net acknowledgment frameworks for substance based web access to paper records. As indicated by the content sort, Handwritten and/or Machine Printed Character Recognition Systems are two fundamental territories of enthusiasm for character acknowledgment field: Machine printed content incorporates the materials, for example, books, daily papers, magazines, records, and different written work units in the feature or still picture. Machine printed characters are basically uniform in stature, width, and pitch expecting the same textual style and size.

These issues for altered textual style, multi text style and Omni textual style character acknowledgment are generally surely known and fathomed with little limitation.

Handwritten text content can be further partitioned into two classes: cursive and hand printed script. Acknowledgment of transcribed characters is a substantially more troublesome issue. Characters are non-uniform and can differ enormously in size and style. Indeed, even characters composed by the same individual can shift significantly. The area of characters is not unsurprising, nor the dispersing between them. In an unconstrained framework, characters may be composed anyplace on the page and may be covered or disjoint. An ordinary acknowledgment framework will oblige an imperative, or included data, about the information being transformed.

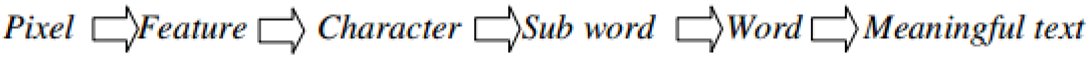
1.3 Optical Character Recognition (OCR)

OCR is the process of converting scanned images of machine printed or the other hand transcribed content into a PC process able arrangement. This is a branch of software engineering that aides in perusing content from paper and making an interpretation of the pictures into a shape that the PC can control. It includes PC programming intended to decipher pictures of typewritten content into machine printed editable content, or to make an interpretation of pictures of characters into a standard encoding plan speaking to them in ASCII or Unicode. OCR is famous for its different application possibilities in banks, post workplaces and guard associations. An OCR framework empowers you to take a book or a magazine article, sustain it specifically into an electronic PC document, and after that alter the record utilizing a word processor. In the event that one outputs a content archive, one may need to utilize optical character acknowledgment (OCR) programming to make an interpretation of picture into content that can be altered. At the point when a scanner first makes a picture from page, picture is put away in PC's memory as a bitmap. A bitmap is a network of spots; one or more bits speak to everyday. The employment of OCR programming is to decipher that show of specks into content that PC can translate as letters and numbers.

Optical Character Recognition (OCR) systems is transforming large amount of documents, either printed alphabet or handwritten into machine encoded text without any transformation, noise, resolution variations and other factors. Optical Character Recognition is the procedure of interpreting pictures of written by hand, typewritten, or printed content into a configuration saw by machines with the end goal of altering, indexing/seeking, and a diminishment away size. There is an awesome need to precisely optically perceive printed materials. A great part of the data is replicated in printed version records. OCR frameworks free this data by changing over the content on paper into electronic structure. Acknowledgment is in this way characterized as the undertaking of content communicated in graphical organization into its typical representation. The pre and post preparing system are easiest method for improving results than different techniques utilized for execution of OCR. Subsequently, a great character acknowledgment approach must dispose of the commotion in the wake of perusing parallel picture information, smooth the picture for better acknowledgment, concentrate includes effectively, prepare the framework and arrange designs. Optical Character Recognition (OCR) is a field of exploration in example acknowledgment, manmade brainpower and machine vision, sign preparing. Optical character acknowledgment (OCR) is generally alluded to as a disconnected from the net character acknowledgment procedure to imply that the framework checks and perceives static pictures of the characters. It alludes to the mechanical or electronic interpretation of pictures of manually written character or printed content into machine code with no variety. Optical Character Recognition (OCR) innovation is utilized to change over data accessible in the printed structure into machine editable electronic content frame through a methodology of picture catch, transforming and acknowledgment.

1.3.1 Stages in OCR

A hierarchical approach for most of the systems would be from pixel to text, as follows:



Pixel ⇨ Feature ⇨ Character ⇨ Sub word ⇨ Word ⇨ Meaningful text

The above various leveled undertakings are assembled in the phases of the character acknowledgment for picture procurement, preprocessing, division, highlight extraction, acknowledgment, and after that at long last post preparing. As indicated in figure 1.2, the methodology of optical character acknowledgment of any script can be comprehensively separated into the distinctive stages.

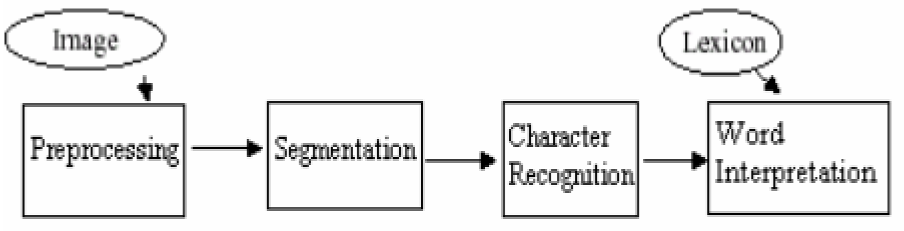


Figure 1.2: Steps in Word Interpretation

1.3.2 Major Steps in character Recognition

Generally there are six major steps in the character recognition system which has been shown in.

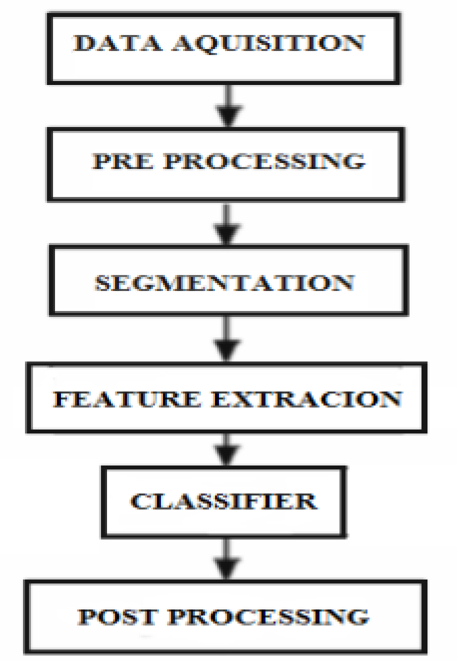


Figure 1.3 Steps in character recognition

A. Data acquisition

The progress in automatic character recognition systems is evolved in two categories according to the mode of data acquisition:

- On-line character recognition system
- Off-line character recognition system

Logged off character acknowledgment catches the information from paper through optical scanners or cams though the on-line acknowledgment frameworks use the digitizers which specifically catches composing with the request of the strokes, velocity, pen- up and pen- down.

B. Pre-processing

1) A progression of operations must be performed amid the preprocessing stages. Preprocessing incorporates all the steps that are important to bring the info information into a structure satisfactory to the up and coming parts of the segments. It incorporates Binarization, limit identification, division, diminishing and picture resizing Binarization. The initial phase in Binarization is to change over a shading picture into a gray scale picture. A gray scale advanced picture is a picture in which every pixel is quantized only the shades of nonpartisan dim, changing from dark at the weakest force to white at the strongest power. To change over any shading to its most estimated dark level, one must get the estimations of essential hues (Red, Green, and Blue). At that point, include 30% of red, 59% of green and 11% of blue esteem together. In any case the scale utilized (0.0 to 1.0 or 0 to 255) the resultant number is the coveted dim worth. The acquired dim picture is then parallels by giving the pixel soul „1‟ for white shade and „0‟ for black shade.

Algorithm Binarization

1. Convert the color image into gray scale image.
2. Find the gray level histogram of input image F(x, y).
3. Find the valley point between two modes and select the point as threshold (T).
4. Binary image b(x,y) is defined as

$$b(x,y) = \begin{cases} 1 & for\ f(x,y) > T \\ 0 & Otherwise \end{cases}$$

At that point locate the dim level histogram of dark scale picture. The dark level histogram is made out of two prevailing modes. One prevailing mode compares to foundation and other is character picture. Select the limit esteem such that the two predominant modes meet at a valley point.

1) Boundary Detection

The limit of character image can be created by stressing locales containing sudden dim light moves and de-underlining districts pretty nearly homogenous force. As such, by examining the picture pixel in both level and vertical heading we can recognize the limit of the characters. In the event of double picture, character pixels are spoken to by dark or '0' esteem and white pixels are by '1' hence by discovery start of '0' quality we can identify the limit of the boundary.

C. Segmentation

Division is a basic stage in OCR system because it impacts the rate of affirmation. Division can be outside and internal. Outside division is the disconnection of distinctive composed work units, for instance, sections, sentences or words. In inward division a photo of collection of characters is broken down into sub-pictures of individual character.

D. Representation or Feature extraction

The highlight extraction step picks and arrangements data which is used by a classifier to achieve the affirmation undertaking. Highlight extraction incorporates addressing a handwriting message by a plan of discriminative highlights. The highlight representation is in light of extraction of particular sorts of information from the photo.

E. Classification

The gathering stage is the decision making bit of the affirmation system. The execution of a classifier relies on upon the way of the highlights. There are various current Classical and sensitive figuring methodology for handwriting unmistakable verification. They are given as:

They are given as:

1) Classical Techniques:

- Template matching
- Statistical techniques
- Structural techniques

2) Soft Computing Techniques

- Neural networks (NNs)
- Fuzzy- logic technique
- Evolutionary computing techniques

F. Post processing

Post-transforming stage is the last phase of the acknowledgment framework. It prints the relating perceived characters in the organized content structure.

1.4 Steps for recognition of a character

Character recognition is not a very easy task. There are several phases required to implement OCR. There are five phases involved as:

1) Image scanning (Digitization),
2) Pre-processing (Binarization),
3) Segmentation,
4) Feature extraction
5) Character recognition

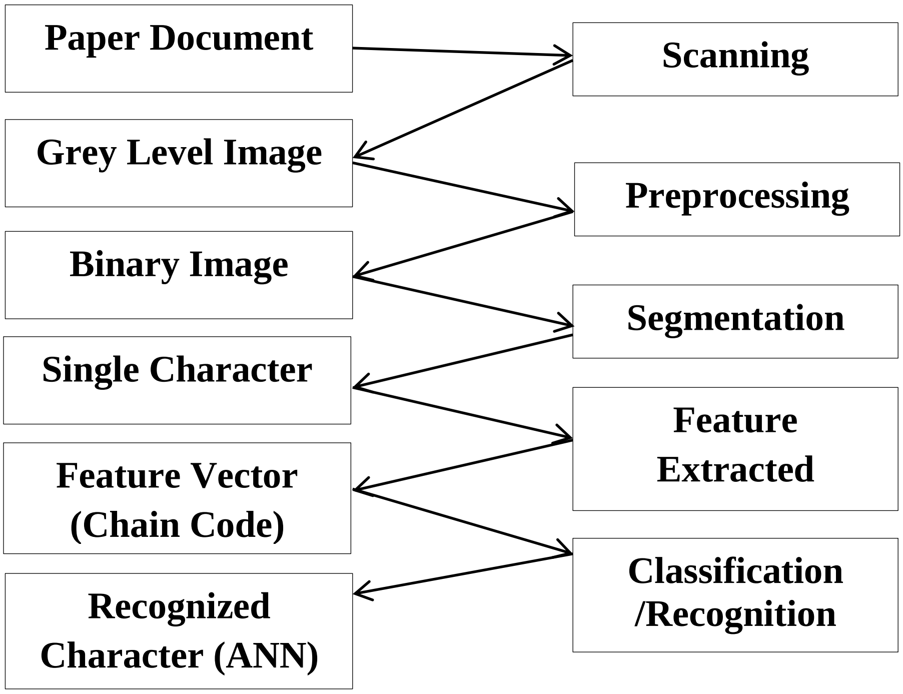


Figure 1.4: Recognition of a character

1.4.1 Processing

Pre-processing is a standout amongst the most crucial venture of image preparing. The data of the pre-processing step is a checked digitized 24- bit bitmap RGB shading picture. Pre-processing is the best approach to change over that filtered RGB picture or shading picture to Grey scale image. Furthermore decipher the Gray scale picture to Binary image.

1.4.2 Segmentation

Division is a standout amongst the most major and essential part for outlining a proficient OCR on the grounds that highlight extraction and acknowledgment methodology relies on upon this stage to build the acknowledgment process effective. The yield of this stage comprises of pictures of individual characters are:

Segmentation process includes the following steps.

- Line Segmentation
- Word Segmentation
- Character Segmentation

1.4.3 Line Segmentation

Concentrate line from a passage by gathering the recurrence for dark pixel in level case. The consequence of line is extraction from the content image.

1.4.4 Word Segmentation

Utilizing vertical output, words are divided by treating the white spaces between two words as a separator. The word division is mostly done by gathering the vertical recurrence of dark pixel.

1.4.5 Character Segmentation

It is the most troublesome and testing branch to manufacture OCR. As Bangla is an inflectional dialect, the ornament of the characters in a word causes numerous eccentricities and makes the division troublesome. The fundamental parts of every last one of characters are situated in the center zone. So the center zone territory is measured as the character division bit. Since matraline interfaces the characters together to frame a word, it is overlooked amid the character division methodology to get them topologically incoherent .There are regularly some white spaces between two characters without matraline in a word. Utilizing vertical output, characters are differentiated by regarding the white spaces as a separator. It is fundamentally done by gathering the vertical recurrence of dark pixel.

1.5 Character Recognition

Conceivable strategies for executing the character acknowledgment calculation were overviewed and at first, a format coordinating procedure was picked. Therefore, an option calculation, based upon the n-tuple system was actualized keeping in mind the end goal to cross-check whether the character acknowledgment execution acquired utilizing the layout coordinating calculation was illustrative for the constrained arrangement of format and test characters accessible.

For hand-printed characters, an unadulterated format coordinating calculation is unrealistic to give great execution due to the wide variety in spatial representation of every character class between diverse essayists. In this work, the format was along these lines altered to make note of the factual variety in character arrangement by misusing thoughts examined in investigations of scattering components for hand printed characters.

1.5.1 Template construction

After every postcode was separated from its envelope picture, it was labelled to show the genuine characters contained in the postcode before partition into the preparation set and test set. The preparation set was then used to collect formats for every character based upon a thickness network Gn, whose directions (x,y) have a worth Gn(x,y), where Gn(x,y) is an evaluation of the likelihood of the related highlight in the preparation examples of class n. Formats were developed basically by applying suitable limits to these thickness networks, thus, for any given Gn and limit m (O S l o %) , the format Tn,,was given as:

Figure 1.4 demonstrates an illustration layout coming about because of applying an edge of 31% to a thickness matrix for the character 3.

More than one format class was characterized for a couple of character classes, to consider basically diverse methods for shaping these Tn,m = ((x,Y) 1 Gn(x,y)>m I characters.

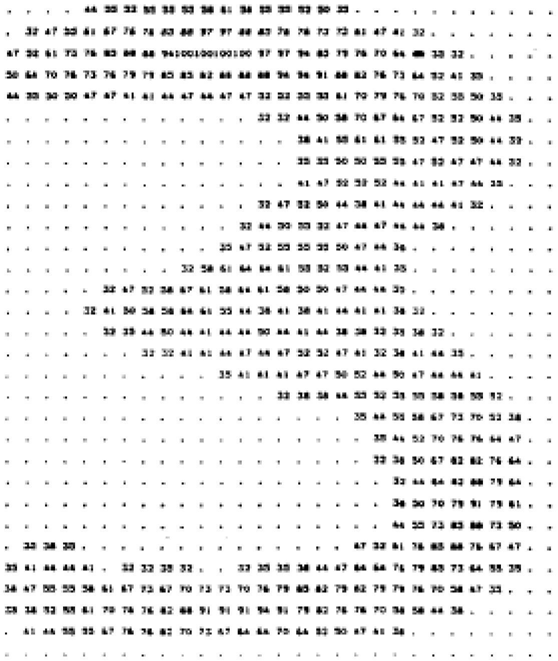


Figure 1.5: Template for the character '3'

1.5.2 Represented graphically for the character

The format coordinating calculation is based after measuring the closeness between a test character, c and a layout, t. There are different definitions for likeness measures: we have decided to utilize a non-metric comparability capacity which basically measures the cosine of the edge between vectors c and t . This likeness measure can be communicated as:

$$s(c,t) = \frac{c^T t}{|c|\,|t|} \qquad (1)$$

where c1 is the transpose vector of c and IcI and It 1 are the sizes of vectors c and t individually. At the point when c and t are both double esteemed with 0,l components, the square of the similitude measure is communicated as:

$$s^2(c,t) = \frac{n_m^2}{n_c \cdot n_t} \qquad (2)$$

where nm is the number of '1' elements in which intersect between the template and the character; is the total number of '1' elements of c; and n is the total number of '1' elements off. Expression (2) is represented graphically in Figure 1.6.

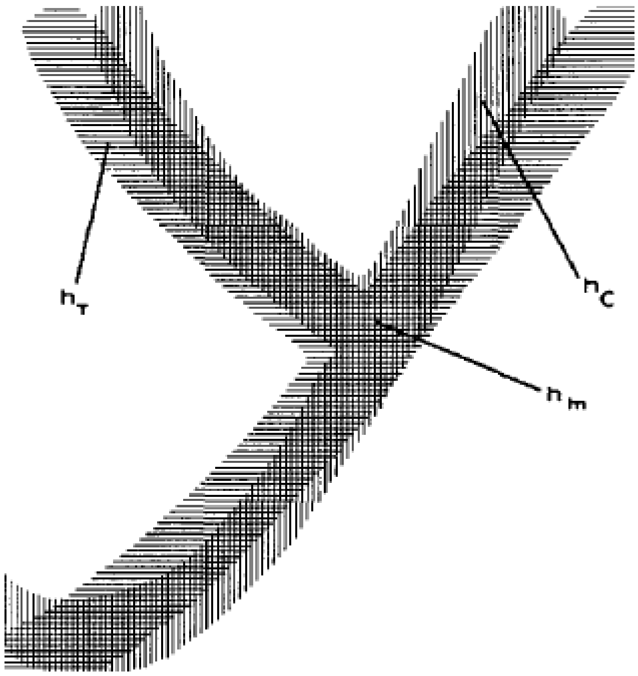


Figure 1.6 s2(c,t) represented graphically for the character 'Y'

On the off chance that c and t are considered as the double representation of the test character and format individually, outflow (2) speaks to the division of pixels of the test character which cross with the layout. To make remittance for the factual dispersion of pixels inside the layout, that is, the likelihood of any specific highlight occurring in the preparation set of the specific class of characters; we then weight (2) to present a coordinating measure as:

On the off chance that c and t are considered as the parallel representation of the test character and layout individually, outflow (2) speaks to the division of pixels of the test character which converge with the format. To make remittance for the measurable dissemination of pixels inside the layout, that is, the likelihood of any specific highlight occurring in the preparation set of the specific class of characters; we then weight (2) to present a coordinating measure as:

$$M_1(c,t) = \frac{n_m^2}{n_c . n_t} \times \frac{\sum p_m}{\sum p_t} \quad (3)$$

This guarantees that M1 is biggest for those test characters which most nearly match the layout.

Another approach to make remittance for the measurable dissemination of pixels inside the format is to substitute the factual weighting element for the variable nm/nt in (2). This gives another coordinating measure as:

$$M_2(c,t) = \frac{n_t}{n_c} \times \frac{\sum p_m}{\sum p_t} \qquad \textbf{(4)}$$

Test results demonstrated some change of acknowledgment rate utilizing M, rather than M2 Since the test character is a twofold example and the format a dim scale example having pt as its components, s2(c,t) could be communicated as:

$$s^2(c,t) = \frac{\left(\sum p_m\right)^2}{n_c \sum p_t^2} \qquad \textbf{(5)}$$

Correlation of (3), (4) and (5) demonstrates that M, and M, are subsequently close estimations to s2. By and by, (5) gives a lower acknowledgment rate than (4) for our present information set, however it may deliver more acceptable results with better layouts.

1.5.3 Image Labelling

Image naming is a two pass calculation which emphasizes through 2 Dimensional double information 8-integration and 4-network name of pixels. 8-integration utilizes North-East, North of the current pixel. 4-integration utilizes North and West of current pixel for marking.

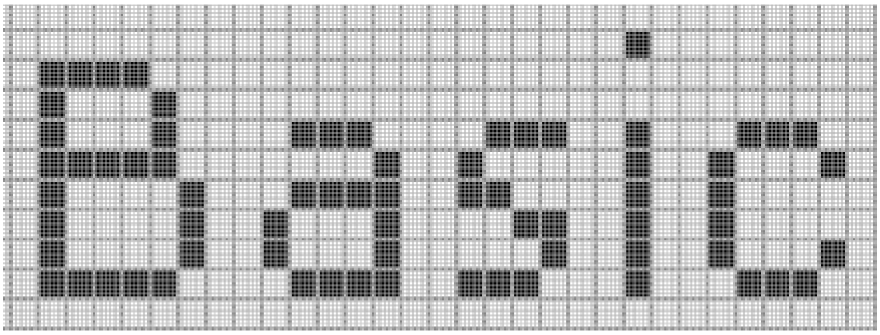


Figure 1.7: Image labelling pattern

Algorithm

Start

{

```
Step 1: Iterate each pixel of data by column and row.
Step 2: Get 8 neighbouring pixels of current pixel.
Step 3: Match the colour of 8 neighbouring pixels with current pixel.
Step 4: Matching colour is added to connected list.
Step 5: If not matching uniquely label the pixel and continue.
Step 6: Otherwise find the connected neighbour with the smallest label and assign to current element.
Step 7: Store the equivalence between the neighbouring labels which are not equal.
End
}
```

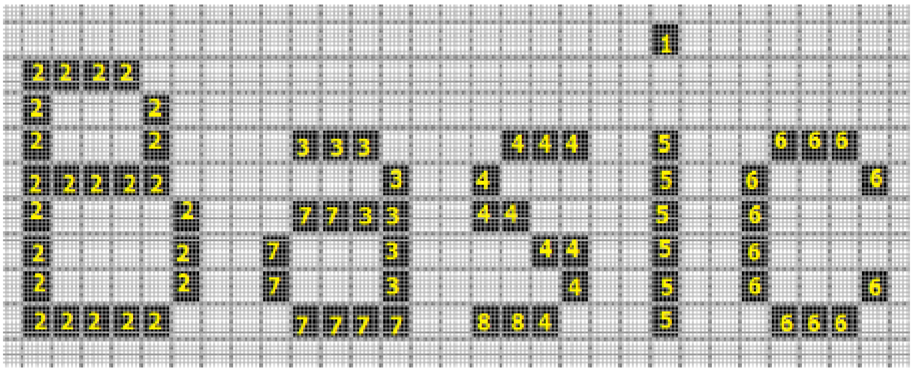


Figure 1.8: Generated Label id's

1.5.4 Finding boundary and Generating X, Y coordinate pixel array

The image demonstrates the whole joined part limit with yellow line. Left organize is given by beginning X arrange and right facilitate is given by consummation X coordinate value.

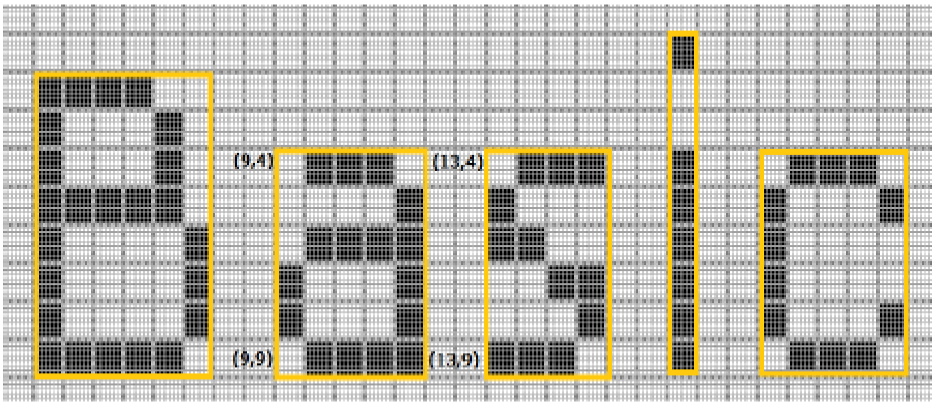


Figure 1.9: Pattern with boundary coordinates

The top record is given by most minimal Y facilitates and base file is given by most astounding Y coordinate. The width is given by right – left organize esteem. The stature is given by base – top direction.

The associated exhibit for the above character is give
A = { (1,0) (2,0) (3,0) (4,1) (4,2) (4,3) (4,4) (4,5) (3,5) (2,5) (1,5) (0,4) (0,3) (1,2) (2,2) (3,2) }
Using the digitized input pattern and this connected array the character is recognized.

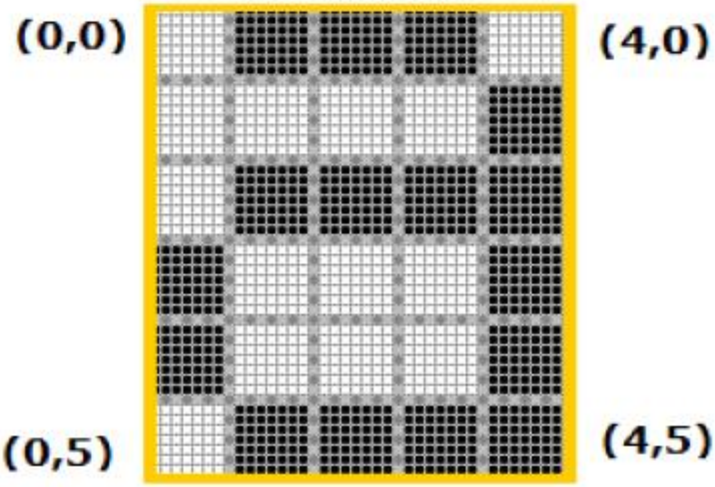


Fig 1.10: Digitized input pattern

1.5.5 Forming Words

The Left X index represent the left most index of the character in the bitmap specified initially in the blog. The Right X index represents the right most X coordinate of the character. When the difference coordinates of current character and previous character is less than 3 pixels then they are joined. This algorithm is quite simple. But you can extend to join words according to grammar in the dictionary.

1.6 Issues Affecting OCR

There are various key issues to consider when watching a printed asset and surveying whether it will create the content asset precision coveted through OCR advances. A percentage of the principle elements are recorded underneath:

1.6.1 Method Used for Scanning

The most obvious component that can enhance OCR exactness is the strategy utilized for examining or determination at which the archive is filtered. 300 dpi is the prescribed best examining determination for OCR exactness. Higher resolutions don't fundamentally bring about better precision and can back off OCR transforming time. Resolutions underneath 300 dpi may influence the quality and exactness of OCR results. Keep in mind that all OCR motors will battle to perceive anything admirably if the determination is beneath 300 dpi and that this is unquestionably the base gauge for checking.

1.6.2 Type of Paper Used

There are different variables identified with the kind of paper utilized that will influence the standard of the filtered picture and these ought to be represented too. The paper on which content shows up is basic to the checked picture standard. On the off chance that the OCR motor can't segregate between the character and the paper foundation clamor then it will be more inclined to distort the character.

1.6.3 Nature of Printing

The way of the printed content in the first may have a huge effect to OCR exactness. Clearly if the content is printed ineffectively or in the event that it was written and characters are broken, blurred or have it defined edges then this will influence the capacity of an OCR motor to perceive examples and separate between comparative formed characters. So the clarity of the printing is an element to consider. A few text styles might likewise have enhanced print clarity over others. Additionally, character sizes of underneath 6 focuses in the first will constrain the precision prone to be accomplished.

1.6.4 Formatting Complications

Varieties in text dimension and sort face may bring about misconception the characters. Broken character and touching character coming from overabundance ink or paper corruptions may not be perceived. Stained and Older archives must be filtered in RGB mode to catch all the picture information, and to augment OCR exactness. These are the couple of reasons which demonstrates that why OCR merchants never guarantee their product to be 100% exact. OCR has been the subject of a huge collection of exploration on the grounds that there are various business applications for this innovation. It can contribute enormously to the progression of a computerization prepare and can enhance the interface in the middle of man and machine in numerous applications. OCR can be utilized for

- Speeding up the information entrance. For information assignment of numerous records, OCR is the most practical and rapidly accessible strategy.
- To decrease information section blunders.
- To decrease the storage room needed by paper reports.

Every year, the electronic stockpiling innovations will free storage room needed to store reports. In the event that these advancements are not utilized then vast space (as cupboards and boxes) is obliged to keep the records. The absolute most noteworthy uses of OCR incorporate the accompanying potential zone where the OCR framework can be helpful:

- Reading guide for the visually impaired
- Automatic content entrance into the PC for desktop production
- Library listing
- Ledgering
- Automatic perusing for sorting of postal mail
- Bank checks and different records
- Document information pressure: from archive picture to ASCII form

- Language transforming the use of transcribed can recognition system can be used in a number of different areas.

1.6.5 To Read Handwritten addresses

In the wake of checking the image of location a piece of the postal wrap, next errand is to allocate a mail piece picture to a conveyance address, by translating written by hand addresses. The location, with the end goal of physical mail conveyance, is comprised of the association name or individual name, essential number which could be a road number or a mail station box number, optional number, for example, a condo or suite number, then took after by road, post office, city, state, nation and pin code. This sort of framework can spare the endeavors of keying the location to framework.

1.6.6 Reading Bank cheque

The system can be utilized for perusing and deciphering the bank checks. This will diminishes the endeavors expected to enter the information composed on the check. Anyhow, bank check acknowledgment shows a few exploration challenges in the range of record examination and acknowledgment. The explanation for this test is a direct result of the shaded foundations and has complex examples to break. To Read Filled structure: The filled structures are for the most part gathered by the diverse association, to gather inputs information from distinctive clients. At times the quantity of these structures is large to the point that just to key in information is a period expending. So the manually written acknowledgment framework can be utilized to decipher the field information of the information shapes.

1.7 Segmentation

Segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). The objective of division is to improve and/or change the representation of a picture into something that is more significant and simpler to examine. Content division is a methodology in which the content picture is isolated into units of examples that appear to shape characters. For text based transforming, after broad system is utilized. To fragment an archive, begin from a given point in the examined picture of report. Portion a region to discover or concentrate the following character. At that point concentrate recognizing properties of character. These characteristics ought to match an individual from a given image set, so next step is to find that image. The procedural grouping, specified above, is rehashed while all extra character pictures are not found. Here one basic inquiry ought to be raised to an analyst, who is taking a shot at division, that what does make a character? The answer is troublesome. The conceivable response to this inquiry can be that a character is an example that takes after one of the images to which the framework is intended to perceive. Be that as it may, to focus such a likeness, the example must be portioned from the archive picture. Accordingly every stage relies on upon the other. Division methodology can comprehensively put in three classes in particular line division, word division and character division. These are examined in the accompanying area.

1.7.1 Line Segmentation

Line segmentation is the process in which the lines are extricated from checked picture. Just lines are separated. Level projection of an archive picture is most generally utilized to concentrate the lines from the record. On the off chance that the lines are very much divided, and are not tilted, the flat projection will have separate tops and valleys, which serve as the separators of the content lines. These valleys are effortlessly recognized and used to focus the area of limits between lines.

1.7.2 Character Segmentation

Character division is the method in which from the fragmented word, characters are removed. Character division is a critical venture of OCR frameworks as it concentrates significant areas for investigation. This step endeavors to break down the picture into classifiable units called character. A poor division procedure produces misrecognition or dismissal. Division procedure is completed after the preprocessing of the picture.

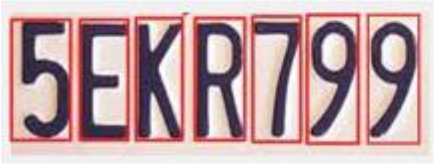


Figure 1.11: Character segmentation

Character segmentation is an operation that looks to decay a picture of a grouping of characters into sub pictures of individual images. Character division is a fundamental essential venture for character acknowledgment in all OCR frameworks. Character division has been an all around examined field in the course of the most recent decade and its principle point was to give individual character to optical character acknowledgment calculations.

1.7.3 Segmentation Strategies

Various methods used can be classified based on the type of text and strategy being followed like the Classical Approach in which segmentations are identified based on character like properties. This process of cutting up the image into meaningful components is also called dissection. Recognition Based Segmentation, in which the framework hunt the picture down segments that match classes in letters in order. Comprehensive Methods, in which the framework tries to perceive words overall, therefore dodging the need to fragment into characters. There are numerous procedures for division, which are blends of one or a greater amount of over three unadulterated ones. Crossover techniques can be spoken to as weighted mixes of these lying at focuses in the mediating space. The three dimensional space speaking to the division methods.

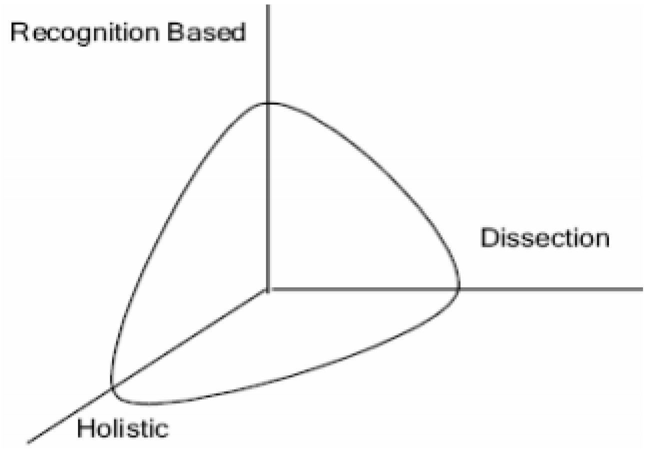


Figure 1.12: Three dimensional space representing the strategies of segmentation

1.8 Optical Character Reader

The OCRs that arrangement with more than one dialect can be of two sorts

1. Multilingual OCRs.
2. Dialect Independent OCRs.

Multilingual OCRs curve prepared for more than one dialect, while the Language Independent OCRs send a procedure that does not get influenced by the dialect of the content being perceived.

The issue of building up an OCR that can manage any dialect is troublesome on the grounds that diverse dialects display distinctive attributes and in this manner summing up their highlights is impractical. Despite the fact that, it is conceivable to make an OCR that can manage more than one dialect yet one conceivable. The current systems for building up the OCRs work over the ligature level acknowledgment that hunt down tile ligatures and concentrate the highlights of every individual ligature, contained in the content. These highlights of ligatures are then utilized for acknowledgment. With ligature level acknowledgment, dialect freedom is difficult to attain to. The recognizer (neural systems for example) ought to be prepared for the characters, contained in all the dialects. Be that as it may, this arrangement gets to be more convoluted for the dialects that permit "cursive-ness" or joining characters. For this case, clearly the recognizer will must be prepared for every single conceivable ligature, instead of simply the character arrangement of the dialect. This results in moderate rate of acknowledgment and lesser exactness.

To conquer these issues for the focus of dialect autonomy, another strategy is proposed. In this strategy, an arrangement of essential geometrical strokes (primitives) is chosen. These strokes ought to be sufficient to speak to all the character arrangement of any dialect. Along these lines a solitary recognizer (neural system and so forth.) would he be able to prepared to perceive those primitive strokes, For the arrangement of the ligatures, a XML record is sent, that contains the structure of the characters of that dialect. This XML record ought to contain just those primitive strokes (and their blends), for which the recognizer is now prepared. One point of interest of this procedure is that recognizer (neural system) must be prepared just once. Another point of interest of this procedure is for cursive dialects in which recognizer require not to be prepared for every word in the lexicon of that dialect. Just the structure of the character set, as far as the essential geometrical strokes and their mixes will be required. OCR framework is utilized to distinguish the writings in the content picture deliberately and changes these records into machine findable and alterable content. In Font acknowledgment, OCR enhances the acknowledgment rate.

1.8.1. OCR Techniques: OCR system contains many techniques to improve the opportunities of successful Font recognition. These techniques are:-

a) **De-Skew:** When the text image is scanned, then it may be aligned in accurate manner, to make the text in perfect manner of horizontal or vertical order, it requires some degree of clock or counter clockwise.

b) **DE speckle:** This technique remove positive and negative spots, smoothing edge.

c) **Binarization:** It is important to font recognition algorithm. A binary image is a technique that converts a color or gray scale image into black or white image. In some cases, this is necessary for the character recognition algorithm. But in some algorithm is performed well on original image, in case this technique is not required.

d) **Line Removal:** Remove the non-glyph lines.

e) **Layout Analysis and Zoning:** It identifies the columns, paragraphs, captions, etc. in particular selected blocks. It is also identify the multiple column layouts and table data.

f) **Line and Word detection:** To detect a word or character shape establishes the baseline.

g) **Script Recognition:** It is very important to identify the script because any document contains many scripts and this situation the script may change at word level.

h) **Character Segmentation:** The character is separated in multiple pieces for particular character recognition in OCR software.

Optical character recognition methods used in our work

In this research methodology, many step used to complete this research. These steps are essential to complete the Font Recognition problem research.

The main Steps are:-

1. Preprocessing
2. Feature extraction
3. Feature Selection
4. Classification
5. Font recognition

1) **Preprocessing:** The preprocessing is used to eliminate the noise and extra things on the image, that help to display the character in more easily way. There are many approaches used to apply the preprocessing: Binarization, Cropping, Gray scale, Noise Reduction, Normalization, Skew Correction, Slant Removal Thinning, and Segmentation etc.

2) **Feature Extraction:** The Feature extraction is defined with two ways:

   i) Local feature: It is related to single letter and based on geometric method

   ii) Global Feature: This feature applied on word, line or paragraph and based on the topological or statistical method.

3) **Feature Selection:** It has three methods:

   i) Filter method: This method have some knowledge of data before the classification.

   ii) Wrapper method: This method has no knowledge about data and dependent on the classifier. It makes the subset of features and after this classification is used.

   iii) Hybrid Method: This method is combination of filter and wrapper and also depends on classifier.

4) **Classification:** The various method used in classification: Weighted Euclidean Distance, Bayesian, Hidden markov model, SVM, Neural networks, Template matching, RBF, and KNN and so on.

Chapter 2
REVIEW OF LITERATURE

G. T. Sutar ,Mr. A.V. Shah (2014) The NPR (Number Plate Recognition) using is a system planned to help in affirmation of number plates of vehicles. This structure is expected with the final objective of the security system. This structure is in light of the photo get ready system. This structure helps in the limits like acknowledgment of the number plates of the vehicles, planning them and using took care of data for further strategies like securing, allowing vehicle to pass or to reject vehicle. NPR is a photo taking care of advancement which uses number (license) plate to perceive the vehicle. The objective is to arrange a successful modified sanction vehicle recognizing confirmation system by using the vehicle number plate. The structure is completed on the section for security control of an extremely restricted zone like military zones or district around top government work environments e.g. Parliament, Supreme Court etc. The made system first gets the vehicle picture. Vehicle number plate district is divided using the photo division as a piece of a photo. Optical character affirmation methodology is used for the character affirmation. The ensuing data is then used to complexity and the records on a database. The system is executed and replicated in Matlab, and it execution is attempted on certified picture. It is seen from the test that the made structure adequately recognizes and see the vehicle number plate on certified pictures. in this paper, the modified vehicle conspicuous evidence system using vehicle tag is presented. The structure use plan of picture changing techniques for perceiving the vehicle from the database set away in the PC. The system is executed in Matlab and it execution is attempted on veritable picture [1].

B.Vani and M. ShyniBeaulah (2014) Optical Character recognition refers to the process of translating the handwritten or printed text into a format that is understood by the machines for the purpose of editing, searching and indexing. The Performance of the current OCR illustrates and explains the actual errors and imaging defects in recognition with illustrated examples. This paper aims to create an application interface for OCR using artificial neural network as a back end to achieve high accurate rate in recognition. The proposed algorithm using neural network concept provides a high accuracy rate in recognition of characters. The proposed approach is implemented and tested on isolated character database consisting of English characters, digits and keyboard special characters. The general throughput of the digit recognizer has been expanded from 10 -12 arrangements every second to 30 arrangements every second. The system of associations what's more, weights acquired by back engendering learning are promptly implementable on business advanced sign transforming equipment. Preparatory results on alphanumeric characters demonstrate that the technique can be reached out to bigger errands. Regardless of the computational Intricacy included, manufactured neural systems offer a few points of interest in example acknowledgment with high exactness in acknowledgment rate with the feeling of copying versatile human knowledge to a little degree [2].

M. Usman Akram, Zabeel Bashir (2013) Optical character acknowledgment is a use of design acknowledgment which naturally distinguishes and perceives the optical characters without human intercession. All the characters are fundamentally made up from three geometric substances, i.e. corners, endings and bifurcations which can be utilized to distinguish diverse characters. In this paper, we show a system for optical character acknowledgment in view of fundamental geometric highlights. The system utilizes an intersection number technique to concentrate highlights from diminished character. The highlight vector for every character comprises of number of corners, endings and bifurcations. The arrangement stage perceives a character by utilizing a straightforward tenet based system. The proposed framework is tried utilizing distinctive tests for every character and the outcomes demonstrate the legitimacy of the proposed calculation. The paper introduced a robotized framework for distinguishing proof of distinctive characters as a utilization of optical character acknowledgment. It connected pre-processing to uproot the clamour and changed over the dim scaled picture into paired by applying an versatile limit. It further connected morphological diminishing operation to facilitate the highlight extraction method. The highlight extraction stage extricated the quantity of corners, closure what's more, bifurcations from diminishing character and utilizing a basic principle based system it distinguishes the character. The proposed framework is invariant to interpretation and turn. The exactness of proposed framework can be enhanced further in the event that we incorporate the unearthly investigation for every character, for example, maximas and minimas [3].

Mohammad Lutf, Yiuming Cheung, C.L. Philip Chen (2013) has created “Arabic Font Recognition Based on Diacritics Features”. This paper presents the alternative approach for the Arabic Font recognition to identify and recognize the font type that is based on the diacritics. Diacritics are like marks and strokes, which are added to the original Arabic Alphabet. So, the Diacritics are the smallest region in the Arabic, which provide the high dimension images with the less cost in present technology. It kept the very important information about the font type. This paper presents the two types of algorithm for diacritics segmentation, such as flood fill and Clustering based algorithms. This method provides the recognition rate of 98.73% as the typical database. This approach is better in case of computation cost and complexity when it is compared with the existing method and it will add with OCR systems in very easy and efficient manner [4].

AkramHajiannezhad and Saeed Mozaffari (2012) proposed "Font Recognition using Variogram Fractal Dimension". This paper presents problem of the font recognition in Arabic, Farsi and English. This approach takes font recognition as a task of texture identification task. In this method the extracted features are not dependent on the text image. This approach is Variogram analysis which is one of the Fractal Dimension Technique. In this method the recognition rate is 95.5%, 96% for Farsi Font, for Arabic Font 96.9%,98.84% and for English fonts 98.21%,99.6% using RBF and KNN classifiers [5].

EhsanMortazaviSenobari, HosseinKhosravi (2012) conducted a survey on "Farsi Font Recognition based on Combination of Wavelet Transform and Sobel-Robert Operator Features". It presents an advanced technique for Farsi Font recognition related to combination of features. Using SRF & Wavelet Transformation method, the features are extracted and combined from textures. These methods are normally different which is having the low correlation for the errors, so that this paper used it for the feature extraction in the better manner. By combining these two methods are easily applicable for the recognition of the text provides to minimize total error and the good result in the implementation. The algorithm which is proposed in this paper is tested on various type of twenty one thousand samples, provide from 10 typical Farsi Fonts. In this proposed technique the characteristics of the font are extracted in clear manner. It provides the recognition rate of 95.56% which is more than SRF and Wavelet transformation [6].

NawrinBinteNawab, M. M. Hassan (2012)

This paper displays an optical BangIa character acknowledgment (OCR) framework utilizing Freeman Chain Code and food forward back engendering neural system which gives much higher execution than some other customary methodologies. After examined the printed BangIa record, preprocessing and division has done. In the highlight extraction stage Freeman Chain Code is utilized. At long last, encourage forward back spread neural system is utilized as a part of grouping of BangIa character. Exhibitions results are displayed in the writing. We portrayed the entire Bangla OCR process orderly saying the calculations needed for every walk. Troublesome character based dialect like Bangia needs no-limit exploration to meet its objective. Absolute work process with novel method for highlight extraction utilizing freeman chain code is compressed in this paper with the goal that it will be steady for analysts on further research field of BangIa OCR [7].

Usha Rani, Balwinder Singh, Ravinder Singh (2012) has developed “Machine printed Punjabi Character Recognition using Morphological operators on Binary images”. It is based on characteristics and there is no requirement of any learning phase and memory. The main benefit of this technique to identify Punjabi fonts in good accuracy. It provide high accuracy on different kind of Punjabi fonts and size that are input scanned image of magazine, newspapers, and any other books [8].

YaghubPoursad, AzamGhorbani, SamanGhouparanloo (2012) presents “Farsi Font and Font size Recognition which is based on analyzing Binarization effect on small document of document images”. It describes an algorithm for Farsi Font and font size recognition. It shows a method that recognizes the fonts and find out the size of font from any document image. It depends on binarization of document and then it gives the effect of binarization of document, it consist the size and shape of dots and broken strokes to recognize text font and font size those are formed in binarization step. For describing this proposed system, a database is consider that include 10*49 text images of seven different fonts in 7 font size that establish in paint software and provide accuracy rate 95.7%. It is also helpful to using in other language such as Urdu and Arabic [9].

YaghoubPourasad, HousehangHassibi, Azam Ghorbani (2011) has developed “Farsi Font Face recognition in letter Level”. It presents the font recognition as the new technique which recognizes the font of the Farsi document in the letter level. It uses the Euclidian distance method between spatial description and gradient value in which font is recognized in every point of boundary of few important Farsi language character in the text image. This system is implemented and developed by the templates that consists a database which include 25 mostly used Farsi Fonts and include five hundred Farsi document images. These are all stored in the typical database. This algorithm will give the average recognition rate of 98.7% in the implementation [10].

Muhammad Tahir Qadri, Muhammad Asif (2009). This paper presents Automatic Number Plate Recognition (ANPR) is a photo taking care of development which uses number (grant) plate to perceive the vehicle. The objective is to diagram a capable modified affirmed vehicle unmistakable evidence system by using the vehicle number plate. The structure is completed on the path for security control of a significantly constrained region like military zones or area around top government work environments e.g. Parliament, Supreme Court et cetera. The made system first distinguishes the vehicle and a while later gets the vehicle picture. Vehicle number plate range is evacuated using the photo division as a piece of a photo. Optical character affirmation system is used for the character affirmation. The consequent data is then used to differentiation and the records on a database so as to compose the specific information like the vehicle's proprietor, spot of enrolment, area, etc. The structure use plan of picture changing routines for recognizing the vehicle from the database set away in the PC. The system is executed in Matlab and it execution is attempted on real pictures. The amusement results exhibit that the system intensely distinguishes and see the vehicle using tag against various helping conditions and can be completed on the entry of an astoundingly constrained extents [11].

DoinaBanciu, ViorelNegru(2007) The current ability to decipher paper archives rapidly and precisely into machine coherent structure utilizing optical character acknowledgment innovation expands the opportunities in archive seeking and putting away, and also the computerized archive preparing. A quick reaction in deciphering substantial accumulations of picture based electronic archives into organized electronic archives is still an issue. The accessibility of a substantial number of preparing units in Grid situations and of free optical character acknowledgment instruments can be misused to deliver a quick interpretation. Taking after this thought, a few tests concerning optical character acknowledgment were performed on a Grid framework also; their outcomes are accounted for in this paper. These outcomes are empowering further advancements of frameworks for record picture investigation utilizing Grid innovations. The following step will comprise in exploring different avenues regarding joined classifiers on computational Grid base. Contemplating the current development of Grid architectures towards administration situated architectures, the outline of Web administrations wrapping existing OCR frameworks or other report investigation devices or joined groups, too of Web administrations composed utilizing Java OCR, is under talk [12].

A.C. Downton, E. Kabir and D. Guillevic(2006)This paper talks about the advancement of an OCR framework for perceiving transcribed and hand printed locations which incorporate a British postcode composed inside character boxes. The framework makes utilization of syntactic data concerning postcodes what's more, a postcode database which interfaces with the character acknowledgment procedure to guarantee that just substantial postcodes are perceived. Proposed substantial postcodes will then be confirmed utilizing semantic highlights of the rest of the location, to deliver a last postcode which both matches the information characters and is good with the rest of the location. Preparatory results are exhibited to show how the starting phases of syntactic verification can enhance character acknowledgment execution. The utilization of exceptionally straightforward and inadequate syntactic data about postcodes has been exhibited to enhance character acknowledgment execution on the outward postcode in general by 54%. A system for consideration of extra syntactic data has been proposed which ought to further enhance execution, furthermore, when consolidated with a present cutting edge character acknowledgment calculation, lead to a superior OCR framework for manually written mail utilizing a hand printed postcode [13].

Ming-Hu Ha, Xue-Dong Tian, Zi-Ru Zhang (2005) has created “Optical Font recognition based on Gabor Filter”. It provides a font recognition method for individual character. This technique consist three steps; in 1st, the fonts are obtained which is based on Gabor filter optimized with genetic algorithm. In second step, a font recognizer used to take the equivalent result by using the fonts and design information of font typesetting. And the final step, the fulfillment after evaluating of font recognition is done relative to the design information. After processing of Font recognition use the typesetting knowledge, which provides the efficient accuracy of font recognition [14].

CHAPTER 3
SCOPE OF THE STUDY

3.1 Problem Formulation

An analyzing the review of literature we found some problems in this thesis work we are providing solution for those problems.

i. The proposed method shall help to improve the limitations that are encountered by the rule based font recognition approach thereby improving its accuracy and modifying its approach.

ii. The feature extraction procedure enforced previously faced misclassification in certain digit; we need to recover these misclassifications.

iii. Fonts in images which have attached noise (shading, distortion) need to be recognized.

iv. More design patterns need to be used as a training data as fonts varies person to person.

CHAPTER 4
OBJECTIVE OF THE STUDY

i. Implement the previous work done on OCR and find out its limitations.

ii. Specifically work on the limitations which are shown on the cluster graph of the previous work.

iii. Enhancing the digit recognition and accuracy of the feature extraction part

iv. Enhancing the character recognition and find out the advantages of the current research over the previous one.

CHAPTER 5
RESEARCH METHODOLOGY

The methodology we are following is for the digit and text identification using Optical Character Recognition has the following steps:

1. Import a unit8 (unsigned integer of 8 bit) class image.
2. Convert it to gray scale (as we do not need the color channels of RGB).
3. The resolution for the gray scale scanning should be around 500 dots per inch.
4. Convert the image into black & white setting a threshold. B&W implies the conversion to logical (binary) class.
5. Use segmentation for the segregation of the digits/characters
6. Feature extraction:
    a. To find out the location of the candidate character/digit, we will be using the morphological connected component of the expected character size.
    b. For matching the template, the character/digit image is taken as a feature vector.
    c. As there are many variations are there in a character, features which are invariant to some characters need to be used such as rotated, mirrored or stretched.
7. Recognition:
    a. Using template matching: similarity or dissimilarity between each template T_1 and character image C_1 is computed. Template T_k , which has the highest similarity measure is identified and has above the specified threshold, is assigned Id as the class label.
8. Verification

CHAPTER 6
RESULT AND DISCUSSION

1. According to our problem formulation, we have discussed about the previous work of our research. It has the following steps and we gained the following results:

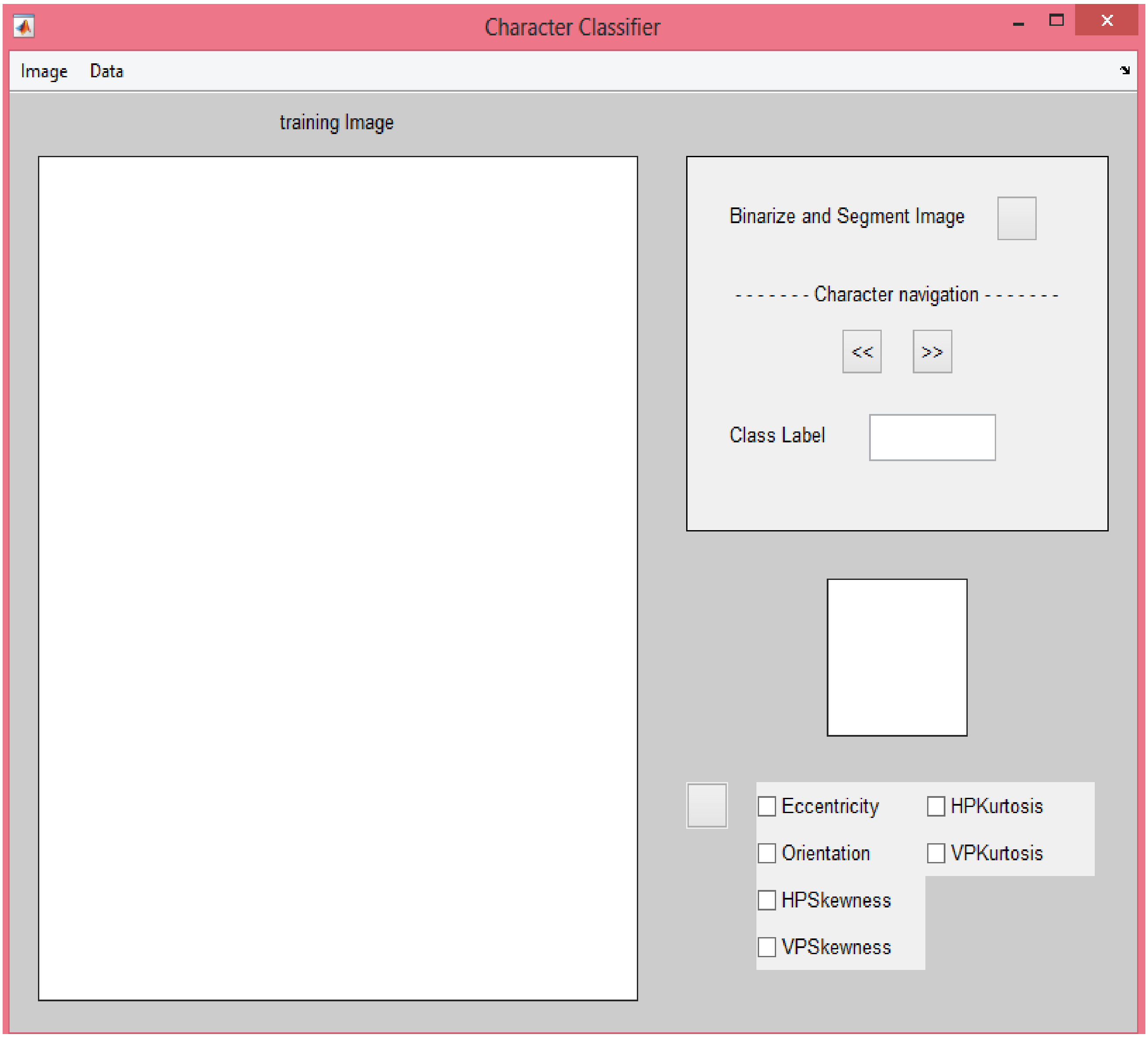
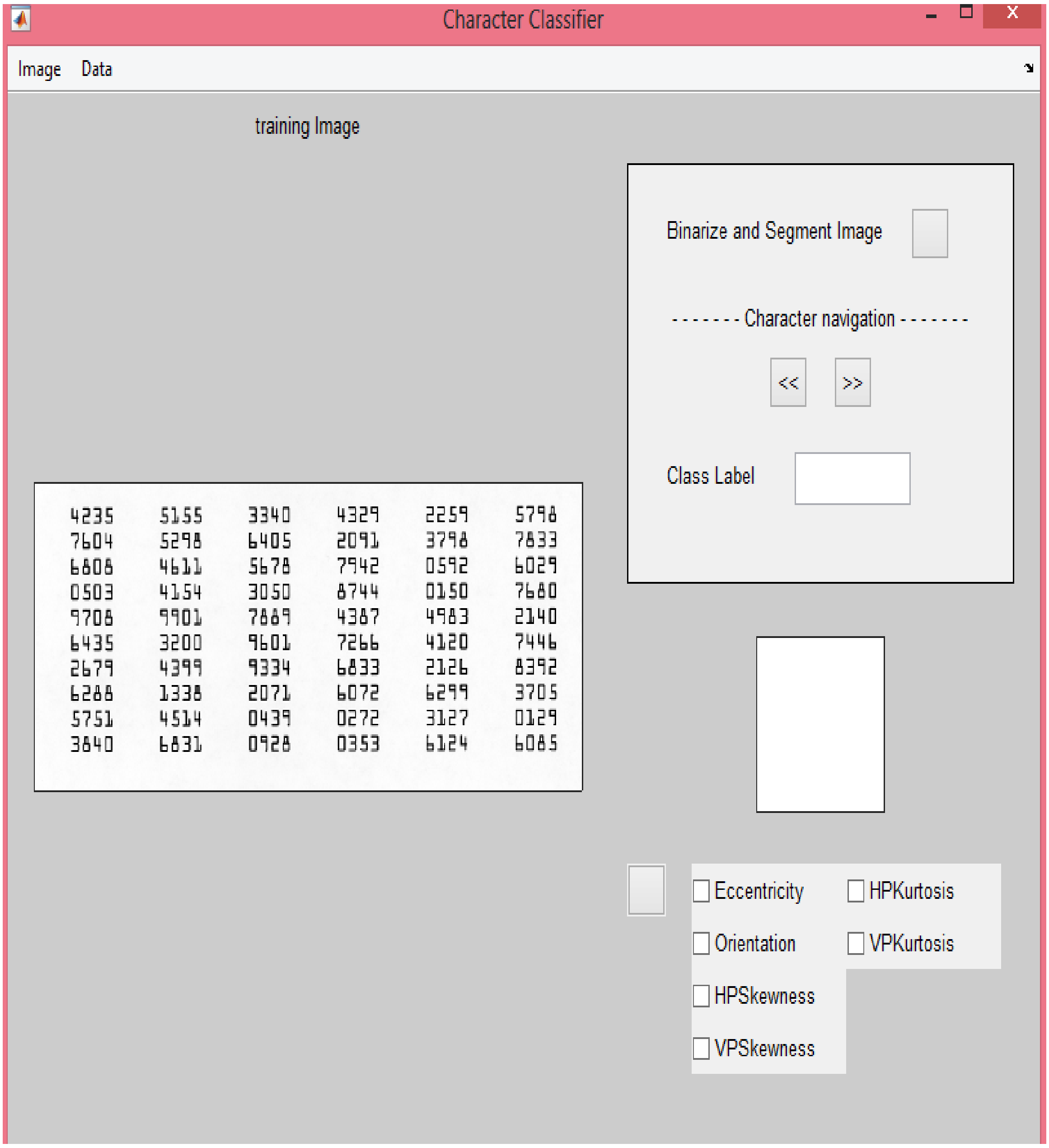


Figure 6.1 The GUI of our previous work.

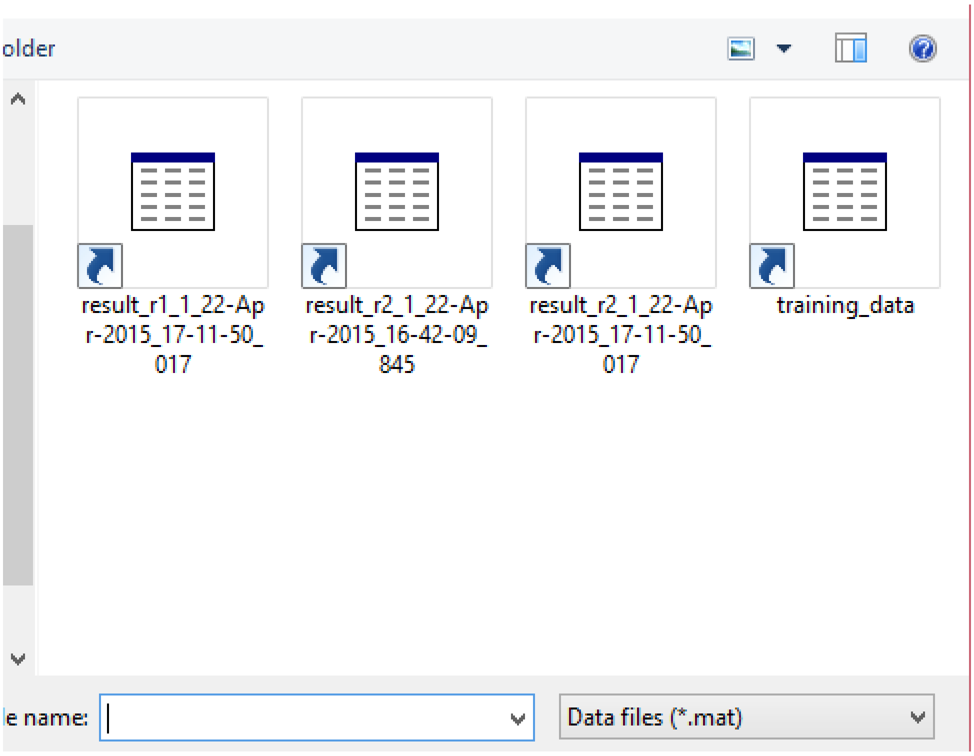
2. We need to import the training images as well as the dataset.



4235	5155	3340	4329	2259	5798
7604	5298	6405	2091	3798	7833
6808	4611	5678	7942	0592	6029
0503	4154	3050	8744	0150	7680
9708	9901	7889	4387	4983	2140
6435	3200	9601	7266	4120	7446
2679	4399	9334	6833	2126	8392
6288	1338	2071	6072	6299	3705
5751	4514	0439	0272	3127	0129
3840	6831	0928	0353	6124	6085

Figure 6.2 Imported training image

3.


older
result_r1_1_22-Apr-2015_17-11-50_017
result_r2_1_22-Apr-2015_16-42-09_845
result_r2_1_22-Apr-2015_17-11-50_017
training_data

le name:

Data files (*.mat)

Figure 6.3 Importing template

4.

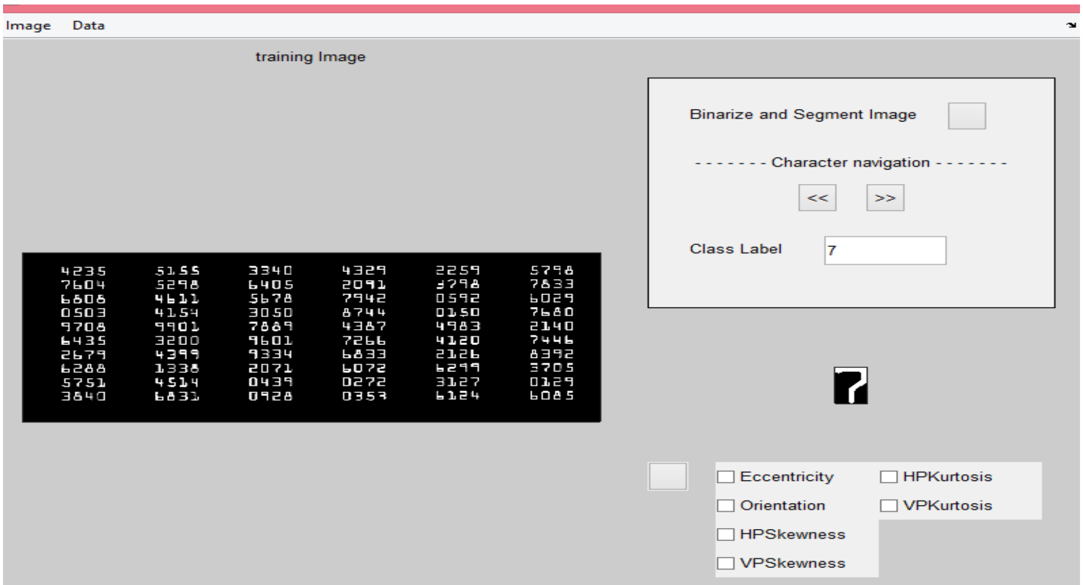


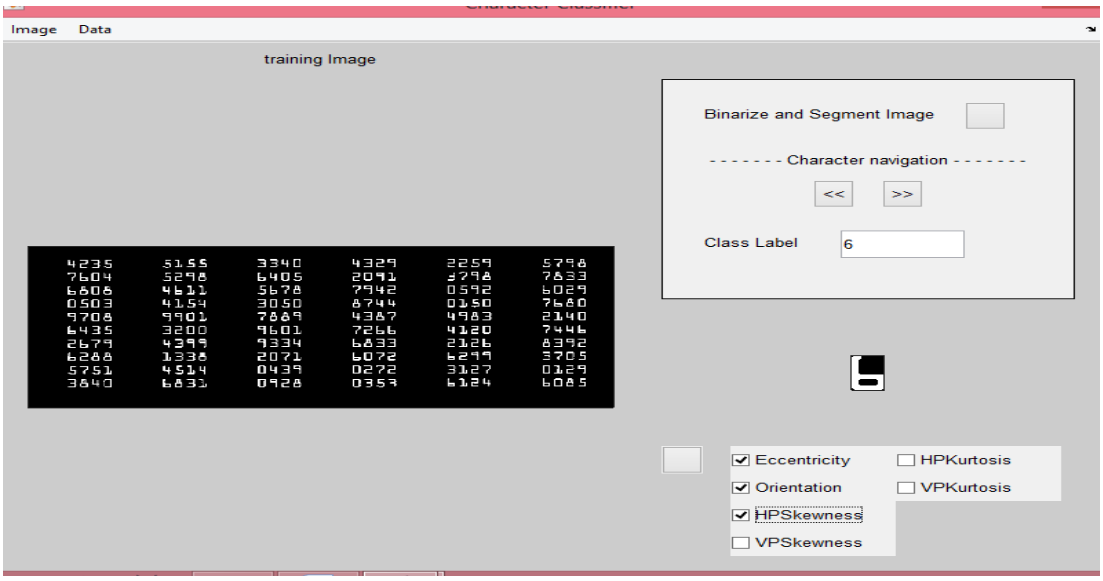
Image
Data
training Image
Binarize and Segment Image
------- Character navigation -------
<<
>>

Class Label 7

☐ Eccentricity
☐ HPKurtosis
☐ Orientation
☐ VPKurtosis
☐ HPSkewness
☐ VPSkewness

Figure 6.4 Binarized segmented image

5.



Class Label 6

☑ Eccentricity
☑ Orientation
☑ HPSkewness
☐ VPSkewness
☐ HPKurtosis
☐ VPKurtosis

Figure 6.5 Select the features for cluster graph

6.

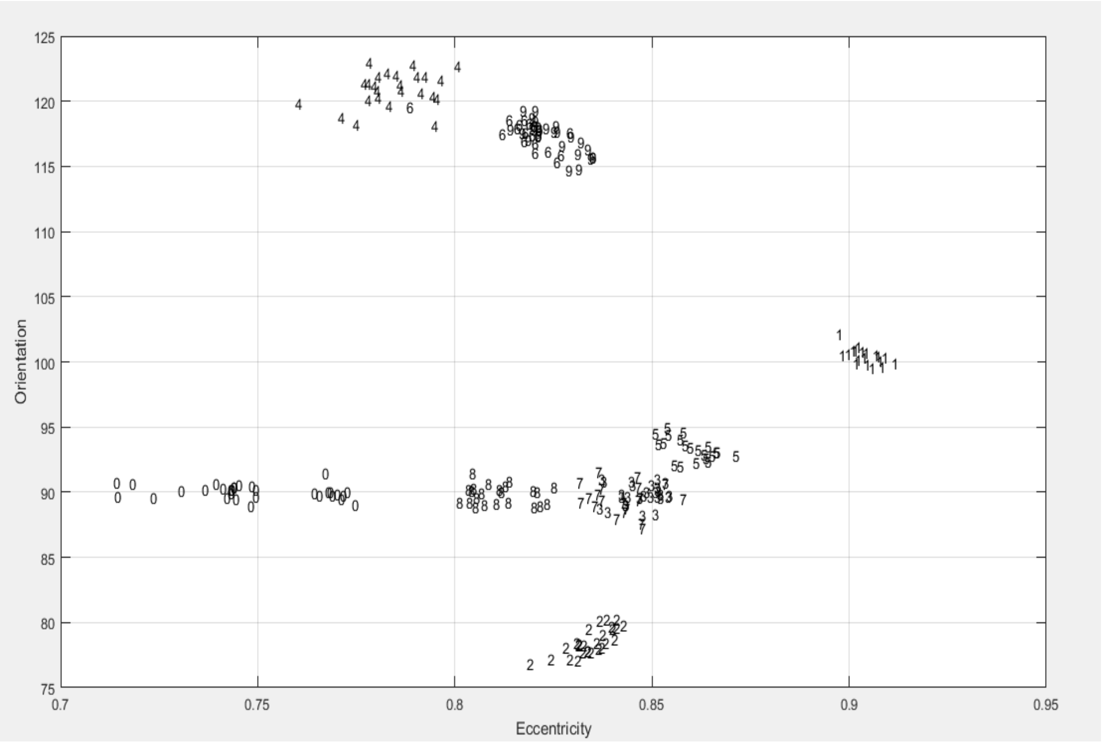


Figure 6.6 Cluster graph

7. Our implementation on Optical Character Recognition

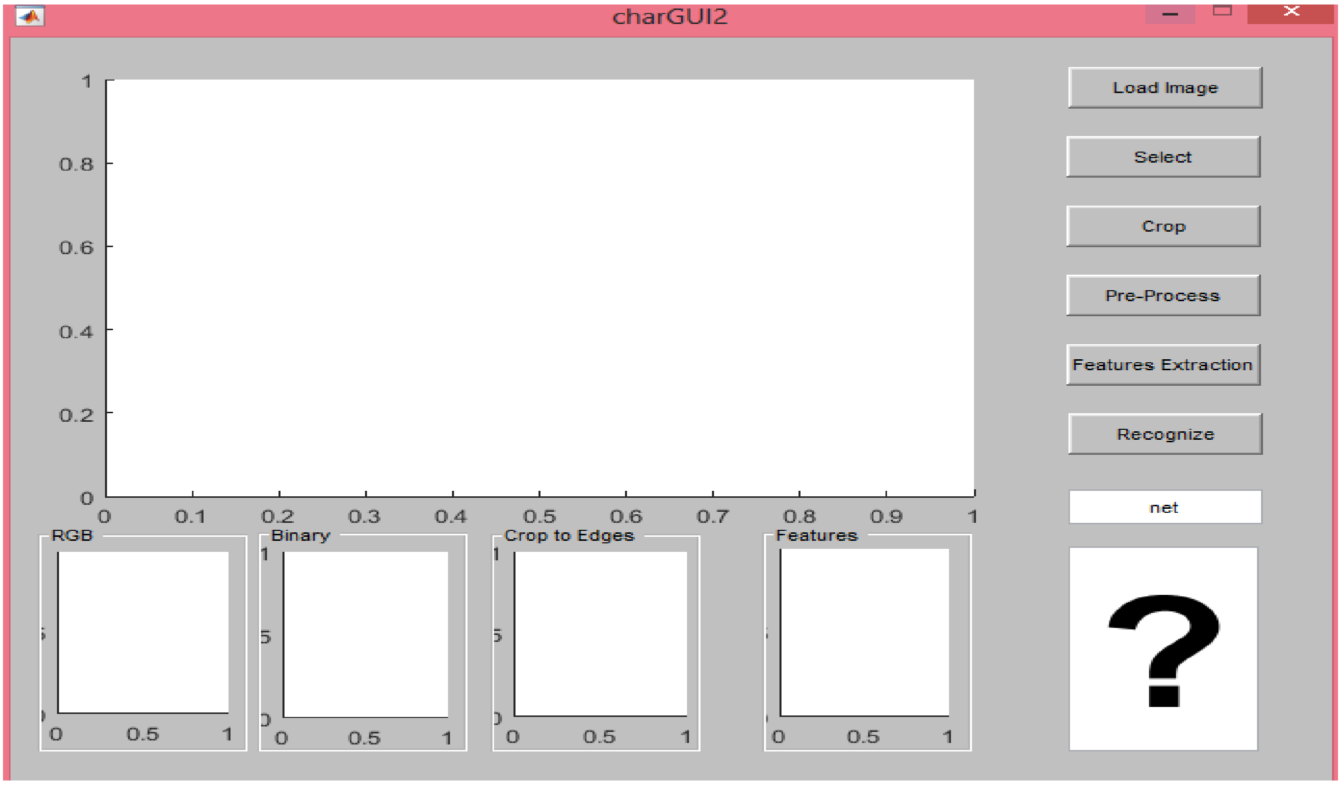


Figure 6.7 GUI

8.

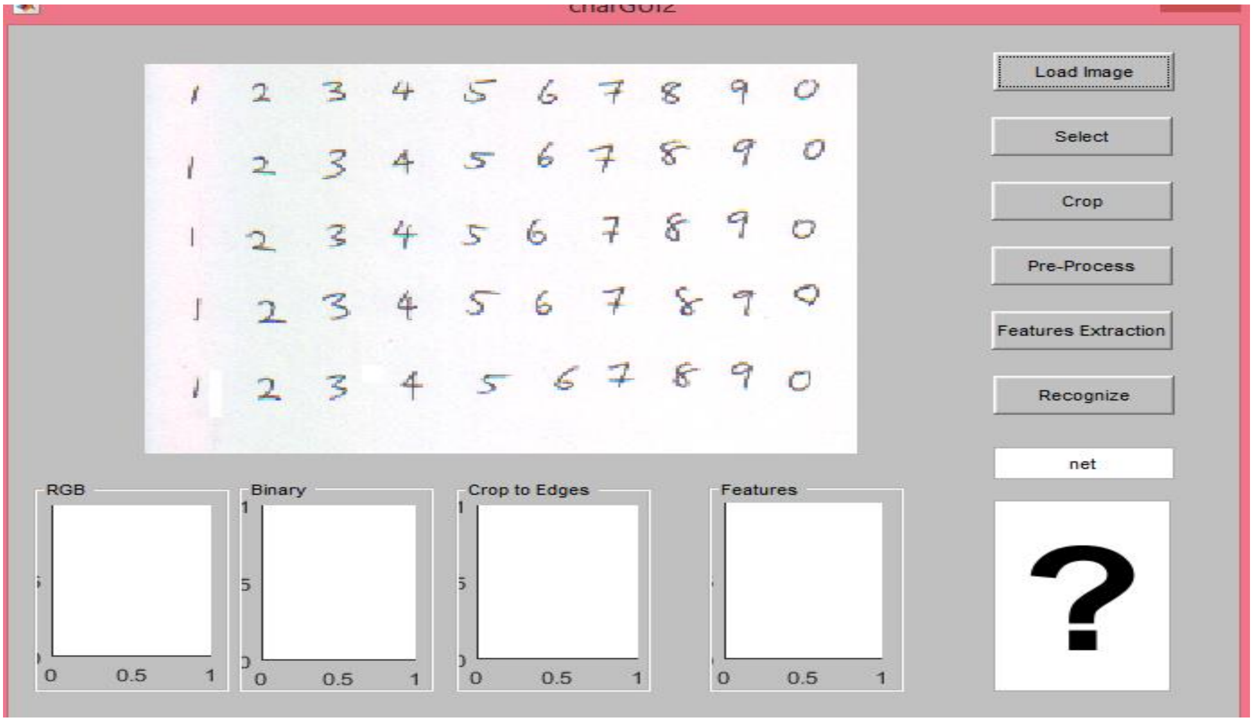


Figure 6.8 Importing image

9.

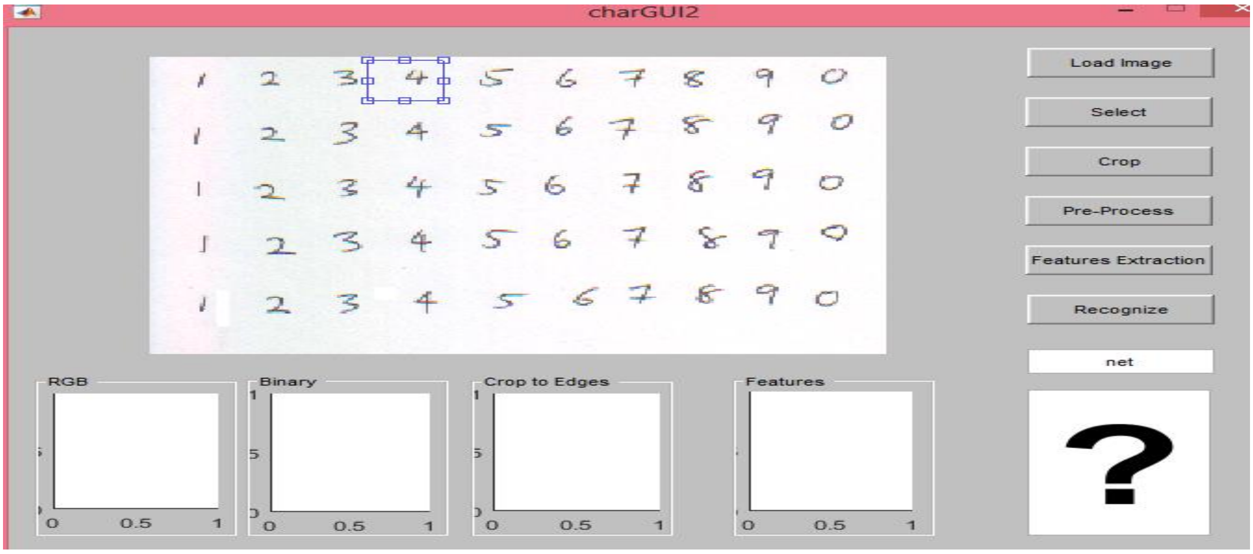


Figure 6.9 Select the specific digit

10.

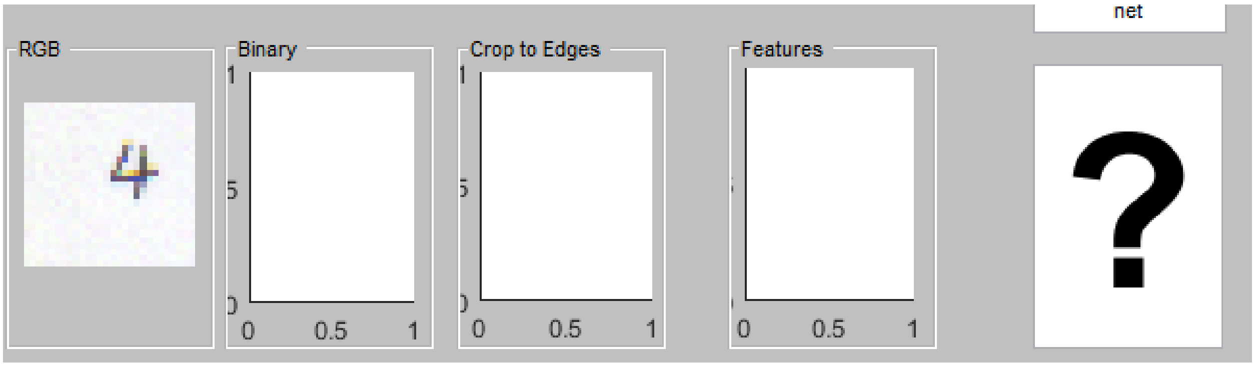


Figure 6.10 Cropping the selected image

11.

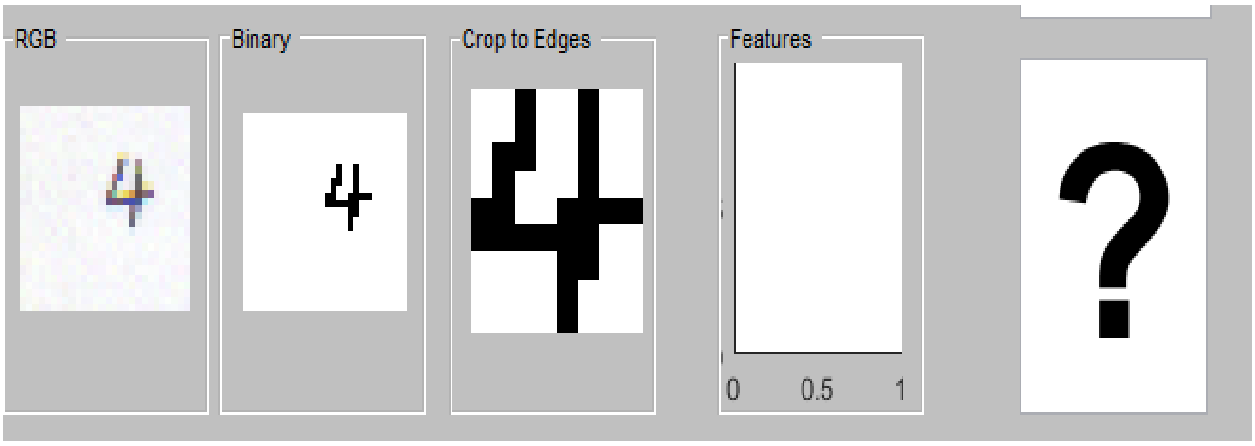


Figure 6.11 Preprocess

12.

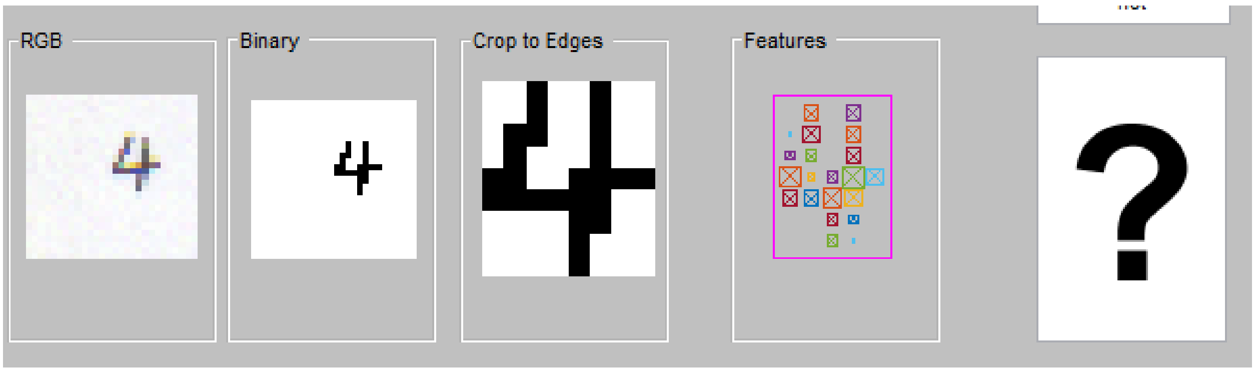


Figure 6.12 Feature extraction

13.

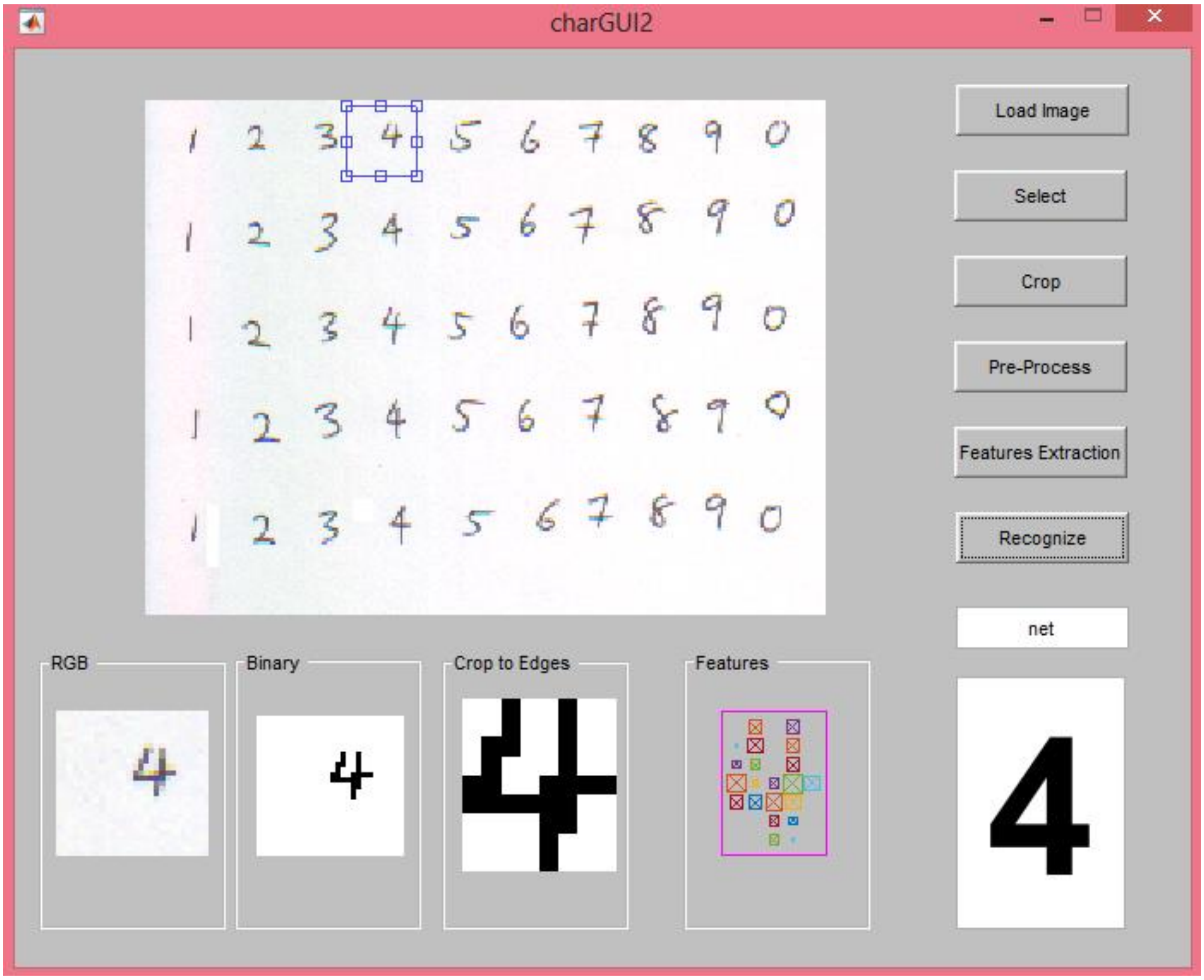
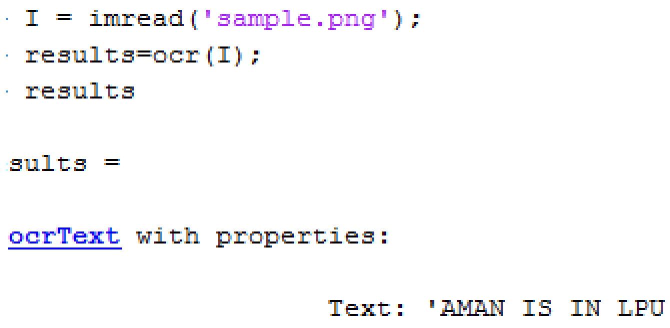


Figure 6.13 Recognition

14.



```
· I = imread('sample.png');
· results=ocr(I);
· results

sults =

ocrText with properties:

                        Text: 'AMAN IS IN LPU
```

Figure 6.14.Recognition of text and digits.

CHAPTER 7
SUMMARY AND CONCLUSION

Our research on Optical character recognition describes the previous work done on OCR as well as implements the same. We have compared the results with the OCR implemented by us which tries to diminish the disadvantages of the previous one. We have reached and optimal results in recognizing the characters. Acknowledgment is frequently trailed by a post-transforming stage. We trust and predict that if post-transforming is done, the exactness will be much higher and after that it could be straightforwardly executed on cell phones. Executing the given framework post-preparing on cell phones is likewise taken as a major aspect of our future work. . Future work of this research also includes working on different feature extraction and segmentation algorithm.

This paper tells about OCR framework for disconnected from the net written by hand character acknowledgment. The frameworks have the capacity to yield brilliant results. Preprocessing procedures utilized as a part of archive pictures as a starting venture in character acknowledgment frameworks were exhibited. The highlight extraction venture of optical character acknowledgment is the most essential. It can be utilized with existing OCR routines, particularly for English content. This framework offers an upper edge by having an advantage i.e. its versatility, i.e. despite the fact that it is designed to perused a predefined arrangement of report configurations, right now English reports, it can be designed to perceive new sorts. The research also describes the previous work done on OCR as well as implements the same. We have compared the results with the OCR implemented by us which tries to diminish the disadvantages of the previous one. We have reached and optimal results in recognizing the characters. Acknowledgment is frequently trailed by a post-transforming stage. We trust and predict that if post-transforming is done, the exactness will be much higher and after that it could be straightforwardly executed on cell phones.

Executing the given framework post-preparing on cell phones is likewise taken as a major aspect of our future work. . Future work of this research also includes working on different feature extraction and segmentation algorithm. Future exploration goes for new applications, for example, online character acknowledgment utilized as a part of cell phones, extraction of content from feature pictures, extraction of data from security archives and handling of authentic records.

CHAPTER 7
LIST OF REFERNCES

1. G. T. Sutar , Mr. A.V. Shah(2014)"Number Plate Recognition Using an Improved Segmentation".

2. *B.Vani, M. ShyniBeaulah(2014)" High accuracy Optical Character Recognition algorithms using learning array of ANN".*

3. Mohammad Lutf, Yiuming Cheung, C.L. Philip Chen (2013) *"Arabic Font Recognition Based on Diacritics Features".*

4. M. Usman Akram, Zabeel Bashir (2013) "Geometric Feature Points Based Optical Character Recognition".

5. QinruQiu, Syracuse University, Syracuse (2013), "*A Parallel Neuromorphic Recognition System and its implementation on a Heterogeneous High-Performance Computing Cluster"*

6. M. U. Akram, S. A. Khan (2013), "Multi-layered Thresholding Based Blood Vessel Segmentation for Screening of Diabetic Retinopathy", Engineering with Computers (EWCO).

7. Akram Hajiannezhad, Saeed Mozaffari (2012) *"Font Recognition using Variogram Fractal Dimension".*

8. Ehsan MortazaviSenobari, Hossein Khosravi (2012) *"Farsi Font Recognition based on Combination of Wavelet Transform and Sobel-Robert Operator Features".*

9. Usha Rani, Balwinder Singh, Ravinder Singh (2012) *"Machine printed Punjabi Character Recognition using Morphological operators on Binary images".*

10. YaghubPoursad, AzamGhorbani, SamanGhouparanloo (2012) *"Farsi Font and Font size Recognition based on analyzing Binarization effect on small document of document images".*

11. NawrinBinteNawab, M. M. Hassan(2012)" *Optical BangIa Character Recognition using Chain-code"*

12. R. Vogt, M. Janeczko, J. LoPorto, and J. Trenkle (2012), "*Neural Network Recognition of Machine-Printed Characters",* Proceedings of the Fifth U.S.P.S. Advanced Technology Conference.

13. Tu, Z., Chen, X., Yuille, A. L., and Zhu, S. C. (2012), *“Image parsing: Unifying segmentation, detection, and recognition”*, International Journal of Computer Vision, Marr Prize Issue.

14. Yaghoub Pourasad, Househang Hassibi, Azam Ghorbani (2011) *“Farsi Font Face recognition in letter Level”.*

15. Holley, Rose "*How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs*" D-Lib Magazine, Retrieved 5 January 2011.

16. Tariq, M. U. Akram (2010)*“An Automated System for Colored Retinal Image Background and Noise Segmentation*”, IEEE Symposium on Industrial Electronics and Applications (ISIEA 2010).

17. Muhammad Tahir Qadri, Muhammad Asif(2009)”*Automatic number plate recognition system for vehicle identification using optical character recognition” .*

18. Dana Petcu, SilviuPanica, DoinaBanciu(2007) ” *Optical Character Recognition on a Grid Infrastructure”.*

19. A.C. Downton, E. Kabir and D. Guillevic (2006)” *Syntactic and contextual post processing of handwritten address for optical character recognition”.*

20. V. Kasmat, and S. Ganesan, (2005)“*An efficient implementation of the Hough transform for detecting vehicle license plates using DSP’s*,” IEEE International Conference on Real-Time Technology and Application Symposium, Chicago, USA.

21. Ming-Hu Ha, Xue-Dong Tian, Zi-Ru Zhang (2005) *“Optical Font recognition based on Gabor Filter”.*

22. Daniel S. Yeung, Zhi-Qiang Li, Xi Zhao Wang (2005) *“Machine Learning and Cybernetics”*

23. Gobina G. Chowdhary (2003) “Natural Language Processing”.

24. J. W. Hsieh, S. H. Yu, and Y. S. Chen (2002)*” Morphology based license plat detection from complex scenes” 1*6th International Conference on Pattern Recognition.

CHAPTER 8

APPENDIX

8.1 List of abbreviations

ANPR	“Automatic Number Plate Recognition”
ASR	“Automatic Speech Recognition”
FD	“Fractal Dimension”
GUI	“Graphical User Interface”
KNN	“K-nearest neighborhood”
OFR	“Optical Font Recognition”
OCR	“Optical Character Reader”
RGB	“Red Green Blue”
SVM	“Support Vector Machine”
WT	“Wavelet Transformation”
XML	Extensible Markup Language