Cleaning and Recognition of Handwritten English Numerals Poor Quality Document consisting of variation in the Brightness

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Technology
in
Computer Science and Applications

Submitted By
Rupali Sharma
(Roll No. 601103017)

Under the supervision of
Dr. Rajiv Kumar
Assistant Professor
(SMCA)

SCHOOL OF MATHEMATICS AND COMPUTER APPLICATIONS
THAPAR UNIVERSITY
PATIALA – 147004

June 2013

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Cleaning and Recognition of Handwritten English Numerals Poor Quality Document consisting of variation in the Brightness*", in partial fulfillment of the requirements for the award of degree of Master of Technology in *Computer Science and Applications* submitted in *School of Mathematics and Computer Applications* Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Rajiv Kumar* and refers other researcher's work which are duly listed in the reference section.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other University.

Rupali Sharma

(Rupali Sharma)

This is to certify that the above statement made by the candidate is correct and true to best of my knowledge.

(Dr. Rajiv Kumar)
Assistant Professor,
SMCA

Countersigned by:

(Dr. Rajesh Kumar)
Professor & Head
School of Mathematics and Computer Applications
Thapar University
Patiala

(Dr S.K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENT

First of all, I would like to express my gratitude to **Dr. Rajiv Kumar, Assistant Professor,** School of Mathematics and Computer Applications, Thapar University, Patiala for introducing me to document image processing and for all his guidance and support. This thesis work was enabled and sustained by his vision and ideas. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps faltered. His patience and support helped me overcome many crisis situations and finish this dissertation.

I am also thankful to our **Head of the Department, Dr. Rajesh Kumar** as well as **PG Coordinator, Mr. Singara Singh, Assistant Professor,** School of Mathematics and Computer Applications, entire faculty and staff of School of Mathematics and Computer Applications and then friends who devoted their valuable time and helped me in all possible ways towards successful completion of this work. I thank all those who have contributed directly or indirectly to this work.

Lastly, I would also like to thank my parents for their years of unyielding love and encourage. They have always wanted the best for me and I admire their determination and sacrifice.

Rupali Sharma

(601103017)

ABSTRACT

It is easy for the human mind to decipher the handwritten characters with accuracy, but for the machines, it is a difficult task. The Optical Character Recognition (OCR) systems have been developed to solve this problem. The output of a scanner or camera captured document is a non editable text image. Though the text is visible but one can neither edit it nor make any change, if required. This provides a basis for the optical character recognition (OCR) theory**.** The overall OCR process consists of three major sub processes like pre processing, segmentation and recognition. Out of these three, the preprocessing process is first phase of OCR system. Preprocessing aims to produce data that is easy for the character recognition systems to operate upon accurately.The proposed approach initiates with the preprocessing of the image through various techniques like average binarization, modified average binarization, skew correction. Binarization is a technique to convert a gray scale into binary (black & white). Sometimes some scanned or camera captured documents often have varying degrees of brightness and require more careful treatment than merely applying average binarization technique. The proposed approach for binarization has solved this problem. The cleaned up image is then, decomposed into lines, words and characters by using the line segmentation, word segmentation and character segmentation techniques respectively. The segmented characters need a proposed Size Normalization technique to adjust the size, then need to be recognized by machine. The segmented characters are analyzed to check the presence of the features like sidebars, closed loops and water reservoirs. Based on these features, a classification structure has been developed, which assigns the segmented character to the predefined class. Most of the approaches for recognition process are based on the neural network technique and its adaptations. These techniques are computationally difficult and require good amount of time to perform the training of the systems in order to provide good results. Thus, there is a need for a simpler approach for the character recognition process. The authors have thus, tried to develop a simple and efficient algorithm for the recognition. All proposed algorithm was applied to several documents and satisfying results have been obtained by the authors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Optical Character recognition (OCR) is used as an umbrella term, which covers all types of machine recognition of characters in various application domains. Character Recognition or Optical character recognition, usually abbreviated to OCR, is electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. Optical character recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer process able format. It involves computer software designed to translate images of text into machine printed editable text, or to translate pictures of characters into a standard encoding scheme representing them in ASCII or Unicode. If anyone scans a text document, one might want to use optical character recognition (OCR) software to translate image into text that can be edited. It is widely used to translate documents, article and books into electronic files, to computerize a record-keeping system in an office, or to publish the text on a website. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning artifacts. In most OCR systems, character recognition performs on individual characters. In OCR process, after capturing the text image, it is passed through various phases. There are generally four phases. These are named as preprocessing, segmentation, recognition and then post processing.

1.1 Significance of OCR and its Usage

Optical character recognition (OCR) software works with our scanner to convert printed characters into digital text, allowing you to search for or edit our document in a word processing program. From speedy searches to saving space, there are numerous benefits to scanning our documents with OCR software:

(i) No more retyping

If we lose or accidentally erase an important digital file, such as a proposal or invoice, but still have a hard copy, you can easily replace it in our digital filing system by using OCR software to scan the paper original or most recent draft.

(ii) Quick digital searches

OCR software converts scanned text into a word processing file, giving we the opportunity to search for specific documents using a keyword or phrase. For example, we could effortlessly search hundreds of invoices and locate a specific name or account in moments, without having to thumb through extensive files.

(iii) Edit text

Once we have scanned our document using OCR, we have the option to edit the text within a word processing program of our choice. Scan items that may need to be updated in the future.

(iv) Save space

Free up storage space by scanning paper documents and hauling the originals off to storage. We can easily turn a filing cabinet worth of information into editable digital files, and create a backup system consisting of a single CD.

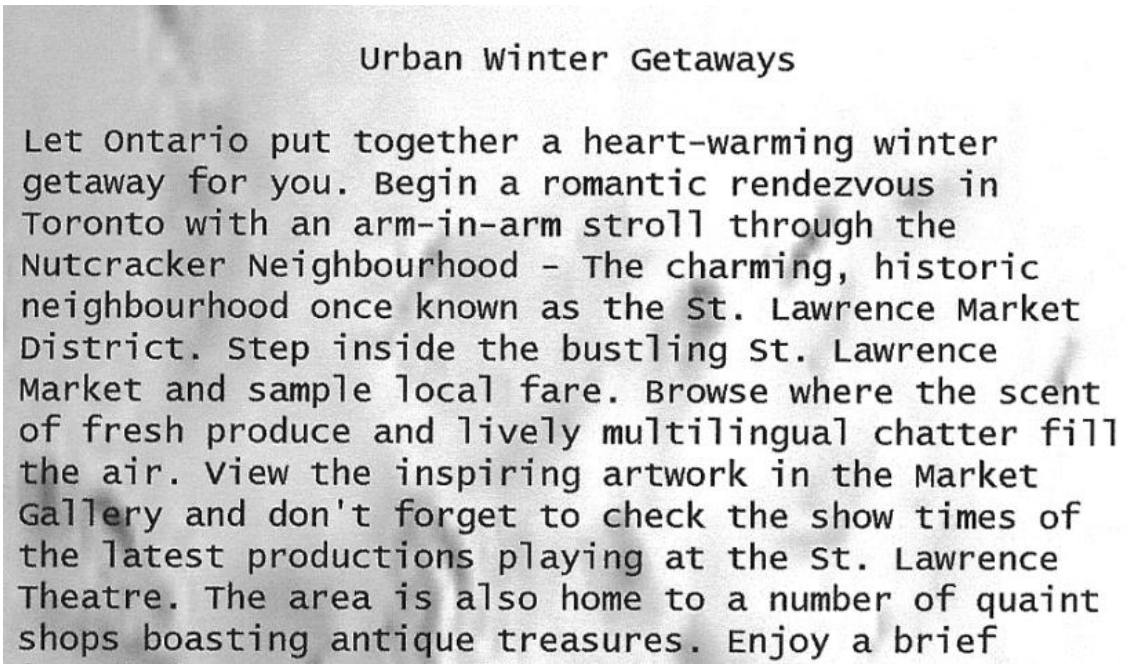
(v) Accessibility

OCR software is a useful Accessibility or Ease of Access tool. Vision impaired PC users can scan books, magazines, incoming faxes, or other documents into word processing programs to be used in conjunction with a computer voice-over utility.

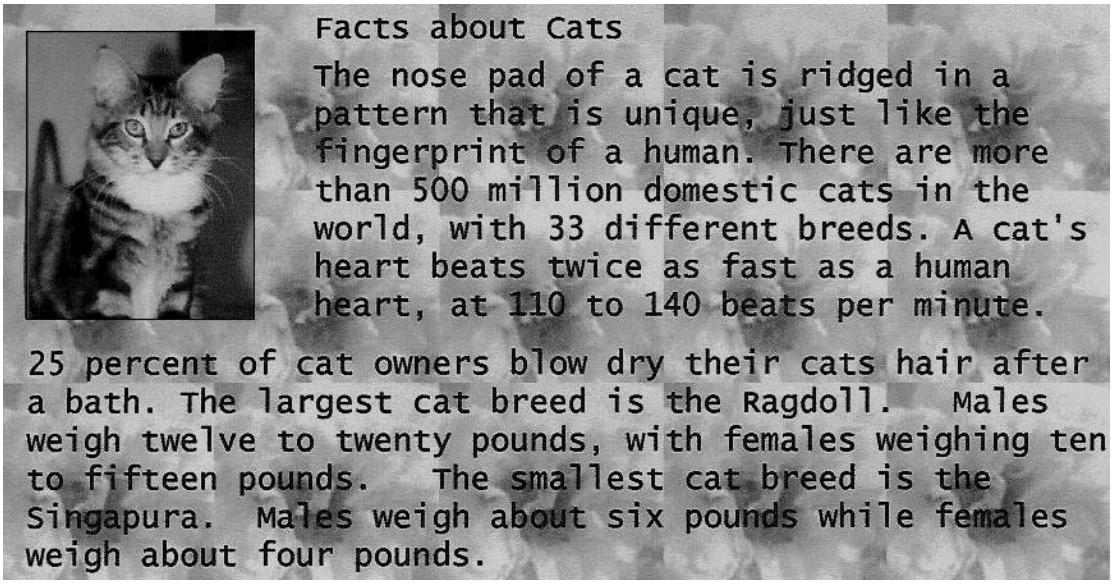
1.2 Factors Affecting OCR Quality

There are a number of factors that affect the accuracy of text recognized through OCR. These factors include: scanner quality, scan resolution, type of printed documents (laser printer or photocopied), paper quality, fonts used in the text, linguistic complexities, and dictionary used.

Watermarks and non-uniform illumination are examples of problems that affect the accuracy of OCR compared to a clean text on a white background. For example, Figure 1.1(a) shows a grey-level document image with poor illumination and Figure 1.1(b) shows a mixed content document image with complex background. Other factors include features of printing such as uniformity, text alignment and arrangement on the page, graphics and picture content



(a) Example of Image with Non-uniform background



(b) Example of image with complex background

Figure 1.1: Images with some Impurities [1]

1.3 History of OCR

Character Recognition or Optical Character Recognition (OCR), is the process of converting scanned images of machine printed or handwritten text (numerals, letters and symbols), into a computer format text (such as ASCII). Methodically, character recognition is a subset of the pattern recognition area. However, it was character recognition that gave the incentives for making pattern recognition analysis matured fields of science. After character recognition these character are converted into speech. Speech is the vocalization form of human communication. Speech communication is more effective medium than text communication medium in many real-world applications.

To replicate the human functions by machines, making the machine able to perform tasks like reading is an ancient dream. The origins of character recognition can actually be found back in 1870 but The David Shepard's invention, which was GISMO - A Robot Reader and Writer, can be said to be the first attempts for the modern version of the OCR. This invention was followed by Jacob Rabinow's prototype machine in 1954, which was able to read upper case and printed output [30].David Shepard's company Intelligent Machines Research (IMR) is said to have been the first to apply optical reading techniques in a commercial situation. They installed a system at Reader's Digest in 1955 [29]. Early systems were very constrained in the sense that they were bound to reading special, artificial fonts, which can be said to be associated with the first generation of OCR. According to Mori, the second generation was characterised by the recognition of the hand printed characters [20].In the initial stages, only numerals could be recognized by these intelligent machines and IBM's 1287 OCR system was the first such system of the second generation machine.

For the old OCR systems, appearing in the middle of the 1980's, the challenge was documents of poor quality and large printed and hand-written character sets. Low cost and high performance were also important objectives, which were helped by the dramatic advances in hardware technology. It has been approximately 47 years since the first automatic handwriting readers were proposed. Since then, astounding progress has been made to enable computers to recognize, interpret and identify machine printed and

handwritten script. Extensive research has been carried but the research into handwriting recognition still continues to be intense at the current stage.

Although more sophisticated OCR-machines started to appear at the market simple OCR devices were still very useful. In the period before the personal computers and laser printers started to dominate the area of text production, typing was a special niche for OCR The uniform print spacing and small number of fonts made simply designed OCR devices very useful. Rough drafts could be created on ordinary typewriters and fed into the computer through an OCR device for final editing. In these way word processors, which were an expensive resource at this time, could support several people and the costs for equipment could be cut. The Next section throws some light on the phases of OCR.

1.4 Various Phases of OCR

The OCR system involves four major phases which are Preprocessing Phase, Segmentation Phase**,** Recognition Phase and Post processing Phase. Here according to Arica [2] the flow chart of the various phases in the figure 1.2 that shows how the scanned image will be converted into computer format, passing through these four phases in OCR.

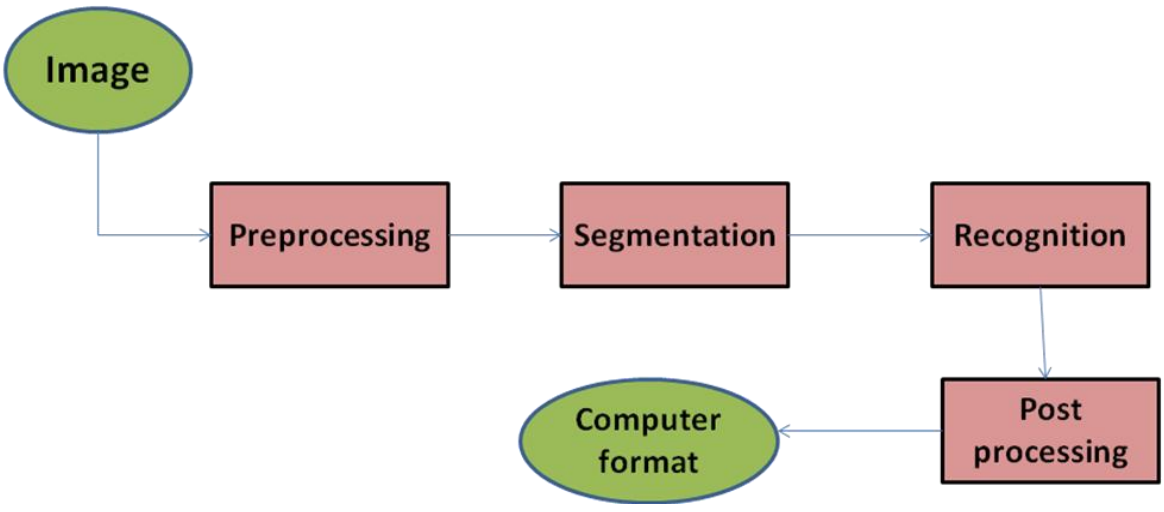


Figure 1.2 Flow Chart of various Phases of OCR

1.4.1 Preprocessing

The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the descriptive phases of character analysis. Preprocessing aims to produce data that are easy for the OCR systems to operate accurately. The main objectives of pre-processing are:-

- **Noise Reduction**

  The noise, introduced by the optical scanning device or the writing instrument, causes disconnected line segments, bumps and gaps in lines, filled loops, etc. The distortion, including local variations, rounding of corners, dilation, and erosion, is also a problem. Prior to the character recognition, it is necessary to eliminate these imperfections.

- **Normalization of the data**

  Normalization methods aim to remove the variations of the writing and obtain standardized data.

- **Reduction in the amount of information to be retained**

  It is well known that classical image compression techniques transform the image from the space domain to domains, which are not suitable for recognition. Compression for character recognition requires space domain techniques for preserving the shape information.

In order to achieve the above objectives, some techniques such as binarization, skew correction, slant normalization, thinning are used in the preprocessing phase.

1.4.2 Segmentation

The preprocessing phase yields a “clean” document in the sense that a sufficient amount of shape information, high compression, and low noise on a normalized image is obtained. The next phase is segmenting the document into its subcomponents. The segmentation process is the most crucial phase. The output of this phase decides the outcome of recognition phase. If this output is right then recognition phase would give the right output otherwise not. In other words we can say that segmentation is an important phase because the extent one can reach in separation of words, lines, or

characters directly affects the recognition rate of the script. One can say that it is one of the decision processes for the OCR. [19] There are two types of segmentation: External segmentation and internal segmentation.

External segmentation which is the isolation of various writing units, such as paragraphs, sentences, or words and internal segmentation, is the isolation of letters, especially in cursively written words.

Character segmentation is all too often ignored in the research community, yet broken and touching characters are responsible for the majority errors in automatic reading of both machine and hand – printed text. Character segmentation is fundamental to character recognition approaches, which rely on isolated characters. It is a critical step because incorrectly segmented characters are not likely to be correctly recognised.

1.4.3 Recognition

OCR process assigns a character image to a class by using a classification algorithm based on the features extracted and the relationships among the features. Since members of a character class are equivalent or similar in as much as they share defining attributes, the measurement of similarity, either explicitly or implicitly, is central to any classifier. Feature extraction is concerned with recovering the defining attributes obscured by imperfect measurements. To represent a character class, either a prototype or a set of samples must be known. The feature selection process attempts to recover the pattern attributes characteristic of each class. Global features, such as the number of holes in the character, the number of concavities in its outer contour, and the relative protrusion of character extremities, and local features, such as the relative positions of line-endings, line crossovers, and corners are commonly used. The classification phase identifies each input character image by considering the detected features [9].

In the statistical classification approaches, character image patterns are represented by points in a multidimensional feature space. Each component of the feature space is a measurement or feature value, which is a random variable reflecting the inherent variability within and between classes. A classifier partitions the feature space into

regions associated with each class, labelling an observed pattern according to the class region into which it falls.

1.4.4 Postprocessing

Until this point, no semantic information is considered during the phases of OCR. It is well known that humans read by context up to 60% for careless handwriting. While preprocesing tries to “clean” the document in a certain sense, it may remove important information, since the context information is not available at this phase. The lack of context information during the segmentation phase may cause even more severe and irreversible errors since it yields meaningless segmentation boundaries. It is clear that if the semantic information were available to a certain extent, it would contribute a lot to the accuracy of the OCR phases. On the other hand, the entire OCR problem is for determining the context of the document image. Therefore, utilization of the context information in the OCR problem Creates a chicken and egg problem. The review of the recent OCR research indicates minor improvements when only shape recognition of the character is considered. Therefore, the incorporation of context and shape information in all the phases of OCR systems is necessary for meaningful improvements in recognition rates. This is done in the post processing phase with a feedback to the early phases of OCR. Spelling checkers are available in some languages, like English, German, and French, etc. String matching algorithms can be used to rank the lexicon words using a distance metric. Statistical information derived from the training data and the syntactic knowledge such as N-grams improves the performance of the matching process. In some applications, the context information confirms the recognition results of the different parts in the document image. In automatic reading of bank checks, the inconsistencies between the legal and the courtesy amount can be detected, and the recognition errors can be potentially corrected. However, the contextual postprocessing suffers from the drawback of making unrecoverable OCR decisions. In addition to the use of a dictionary, a well-developed lexicon and a set of orthographic rules contribute a great deal to the recognition rates in word and sentence levels. In word level, lexicon-driven matching approaches avoid making unrecoverable decisions at the postprocessing phase by bringing the context information earlier in the segmentation and recognition phases. A

lexicon of words with a knowledge based is during or after recognition phase for verification and improvement purpose.

1.5 Overview of Preprocessing Phase

Now we are going to discuss Pre-processing phase in detail. We will discuss here some pre-processing techniques. Before understanding that how preprocessing is done, we should understand the significance of preprocessing Phase in OCR.

1.5.1 Significance of Preprocessing Phase in OCR

The importance of the preprocessing stage of a character recognition system lies in its ability to remedy some of the problems that may occur due to some of the factors presented in above section. Thus, the use of preprocessing techniques may enhance a document image and then preparing it for the next stage in a character recognition system. In order to achieve higher recognition rates, it is essential to have an effective preprocessing stage; therefore, using effective preprocessing algorithms makes the OCR system more robust mainly through accurate image enhancement, noise removal, image thresholding, skew detection/correction, page segmentation, character segmentation, character normalization, slant normalization, thinning, and binarization.

1.6 Preprocessing Techniques

There are so many techniques to remove the impurities from the scanned image. Impurity in a scanned image may arrive due to several reasons such as bad quality of paper or bad quality of camera or due to bad quality of scanner. These impurities can be removed using some techniques. Five techniques are going to be discussed here:

(i) Binarization

(ii) Skew Correction

(iii) Slant Normalization

(iv) Thinning

(v) Size Normalization

1.6.1 Binarization of an Image

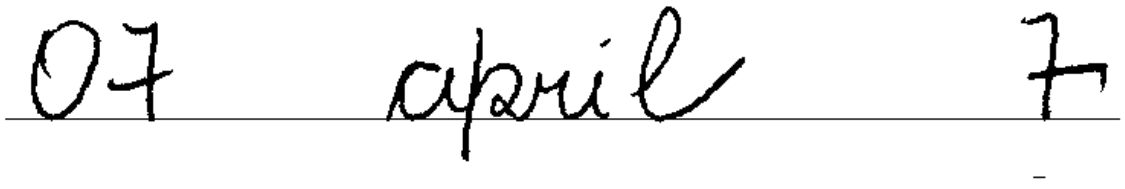
A **binary image** is a digital image that has only two possible values for each pixel. Typically the two colours used for a binary image are black and white. binarization is used as a pre-processor before OCR. In fact, most OCR packages on the market work only on binary (black & white) images. [27] Image binarization converts an image of up to 256 gray levels to a black and white image or on other words we can say that image binarization converts a gray level image into binary image.

Figure 1.3: Conversion of Gray scale image to Binary image

1.6.2 Skew Correction

Due to inaccuracies in the scanning process and writing style, the writing may be slightly tilted or curved within the image. This can hurt the effectiveness of later algorithms and therefore it should be detected and corrected. By nature, handwriting is very unsteady in shape and quality of tracing. The fact that research in this domain has been going on for about thirty years evidences that an all-encompassing solution is still to be found. Handwriting location and recognition phases depend to a large extent on its disposition and more especially on its skew. A lot of words present an unknown arbitrary skew that can generate mistakes in the extraction of these lines and failure of the recognition process. Handwriting skew correction becomes necessary. It is used both in handwritten and printed contexts. After skew detection, the character or word is translated to the origin, rotated, or stretched until the baseline is horizontal and retranslated back into the display screen space.

Original image



After Skew Correction

Figure1.4: Example of Skew Correction [21]

1.6.3 Slant Normalization

One of the measurable factors of different handwriting styles is the slant angle between longest stroke in a word and the vertical direction. Slant normalization is used to

normalize all characters to a standard form. The most common method for slant estimation is the calculation of the average angle of near-vertical elements. Vertical line elements from contours are extracted by tracing chain code components. Coordinates of the start and end points of each line element provide the slant angle. Another study uses an approach in which projection profiles are computed for a number of angles away from the vertical direction. In particular, the slanted characters slope either from right to left or vice versa. Moreover, deferent deviations may appear not only within a text but also within a single word. Some examples illustrating these cases are shown in Fig 1.5.

(a) a word slanted to right

(b) a word slanted to left

(c) a variant-slanted word

Figure 1.5: Examples of slanted word [17]

1.6.4 Thinning

The purpose of thinning is to reduce the image components to their essential information so that further analysis and recognition are facilitated [33]. For instance, the same words can be handwritten with different pens giving different stroke thicknesses, but the literal information of the words is the same. It provides a tremendous reduction in data size, thinning extracts the shape and context information of the characters. It refers to the

process of reducing the width of a line like object from many pixels wide to just single pixel. This process can remove irregularities in letters and in turn, makes the recognition algorithm simpler because they only have to operate on a character stroke, which is only one pixel wide. The basic iterative thinning operation is to examine each pixel in an image within the context of its neighborhood region of at least 3 x 3 pixels and to peel the region boundaries, one pixel layer at a time, until the regions have been reduced to thin lines.

Figure 1.6 Example of Thinning [22]

1.6.5 Size Normalization

Handwriting is subject to a large amount of variability. Normalization ideally aims at the complete elimination of variability. It is used to adjust the character size to a certain standard. Normalization can be done according to the OCR need.

Figure 1.7: Size Normalization of Character **"e"** [2]

1.7 Segmentation

The preprocessing stage yields a clean document in the sense that a sufficient amount of shape information, high compression, and low noise on a normalized image is obtained. The next stage is segmenting the document into its subcomponents. Segmentation is an important stage because the extent one can reach in separation of words, lines, or characters directly affects the recognition rate of the OCR.

Segmentation is the most challenging part of developing recognition software. Some systems force segmentation by providing individual boxes to write in instead of one flowing input section. It is seen that the segmentation decision is interdependent with local decisions regarding shape similarity and with global decisions regarding contextual acceptability. It is a process that seeks to decompose an image of a sequence of characters into sub images of individual symbols by segmenting lines and words. There are two types of segmentation namely, external and internal segmentation as per Arica and Vural [2] are explained in the next section.

1.7.1 External Segmentation

It is the isolation of various writing units, such as paragraphs, sentences, or words. It is the most OCR critical part of the document analysis, which is a necessary step prior to the offline character recognition. Although document analysis is a relatively different research area with its own methodologies and techniques, segmenting the document image into text and non text regions is an integral part of the optical character recognition software.

1.7.2 Internal Segmentation

It is the isolation of letters, especially in cursively written words. Although the methods have progressed remarkably in the last decade and a variety of techniques have emerged, segmentation of cursive OCR script into letters is still an unsolved problem. The character segmentation strategies namely, explicit segmentation, implicit segmentation and mixed strategies are explained here:-

- **Explicit Segmentation**

  In this strategy, the segments are identified based on “character like” properties. The process of cutting up the image into meaningful components is given a special name, dissection.

  Dissection is a process that analyzes an image without using a specific class of shape information. The OCR for good segmentation is the agreement of general properties of the segments with those expected for valid characters.

- **Implicit Segmentation**

  This segmentation strategy is based on recognition. It searches the image for components that match predefined classes. Segmentation is performed by the use of recognition confidence, including syntactic or semantic correctness of the overall result. In this approach, two classes of methods can be employed, which are methods that make some search process and the methods that segment a feature representation of the image.

- **Mixed Strategies**

  They combine explicit and implicit segmentation in a hybrid way. A dissection algorithm is applied to the image, but the intent is to over segment, *i.e.*, to cut the image in sufficiently many places that the correct segmentation boundaries are included among the cuts made. Once this is assured, the optimal segmentation is sought by evaluation of subsets of the cuts made. Each subset implies a segmentation hypothesis, and classification is brought to bear to evaluate the different hypothesis and choose the most promising segmentation.

  The next section elaborates upon the hierarchy of the segmentation of the document image.

1.7.3 Segmentation Hierarchy

The segmentation hierarchy of the document can be described based on the structure of the document. A document generally consists of several pages. A page usually contains

several lines of text, which are, in turn made up of words. The words can be further seen as a grouping of characters. Based on this structure, the segmentation hierarchy can be defined as shown here:-

- **Page Segmentation**

  Page segmentation can be defined as the process of extracting pages from a document.

- **Line Segmentation**

  Line segmentation can be defined as the process of extracting lines from a page.

- **Word Segmentation**

  Word segmentation can be defined as the process of extracting words from a line.

- **Character Segmentation**

  Character segmentation can be defined as the process of extracting characters from a word. The various strategies that can be applied to perform the segmentation of the document according to the above explained hierarchy. The strategies for segmentation are dealt with, in the next section.

1.7.4 Segmentation Strategies

According to Casey and Lecolinet [9], the segmentation strategies, which can be categorized as classical approach, recognition based segmentation approach and holistic approach are discussed here:-

- **Classical Approach**

  In classical approach, segments are identified based on "character like" properties such as height, width, separation from neighboring components, disposition along a baseline, etc. This technique cuts image into a sequence of sub images, which are called dissections.

  The criterion for good segmentation is the agreement of general properties of the segments obtained with those expected for valid characters. Examples of such properties are height, width, separation from neighboring components, disposition along a baseline, etc. This process of cutting up the image into meaningful components is given a special name, dissection. Under dissection, there are various techniques like White Space and Pitch, Projection Analysis, Bounding Box analysis.

- **Recognition Based Segmentation Approach**

  After an image has been segmented into regions, it is ready to enter the next level that is, the feature extraction stage. The result of the image acquisition, preprocessing, segmentation and character fragmentation is a matrix of numbers that represents a character fragment in some way. In the general case, however, the matching of these numbers to a template may be too time consuming and not flexible enough. Therefore, feature extraction is needed. It assigns an input character to one of many pre specified classes, which are based on the extracted features and their analysis.

  Recognition based segmentation, in which, the system searches the image, for those components that match the classes in its alphabet. Many people think that the existence of reliable features to distinguish boundaries in all fonts from interior regions is arguable and the open loop approaches, segmentation to recognition render errors irrecoverable. Therefore, character segmentation should be closely coupled with character recognition.

  Structural features of each character fragment are extracted in this method. The recognition based segmentation techniques are sliding window method, closed loop segmentation and recognition, multiple hypothesis scheme.

- **Holistic Approach**

  There are many applications that require the recognition of unconstrained handwritten words. A word can be either purely numeric as in the case of a ZIP Code, or purely alphabetic as in the case of US state abbreviations, or mixed as in the number of an apartment. In general, a character string recognizer has many applications. The applications include, but are not limited to, reading bank cheques, reading tax forms and interpretation of postal addresses. The task becomes particularly challenging when adjacent characters in a character string are touching. Unlike purely alphabetic strings where joining of the characters is natural and takes place by means of ligatures, the joining of numerals in a numeric word and the uppercase characters in an abbreviation are accidental. The various ways in which two digits can touch are categorized.

  In holistic method the system seeks to recognize words as a whole, thus, avoiding the need to segment into characters. Holistic methods in essence revert to the classical

approach with words as the alphabet to be read. This method performs feature extraction in the first step, then global recognition by comparing the representation of the unknown word with those of the references stored in the lexicon. Consequently, this method uses the classical approach.

One of the advantages of the holistic approach is that no distinction needs to be made among touching and non touching pairs. A segmentation based method must necessarily make such a determination. While the holistic method applies equally to both touching and non-touching characters, its advantage over traditional segmentation based methods is more pronounced among the touching character pairs.

The explained segmentation strategies can be represented in the 3D space, as shown below:-

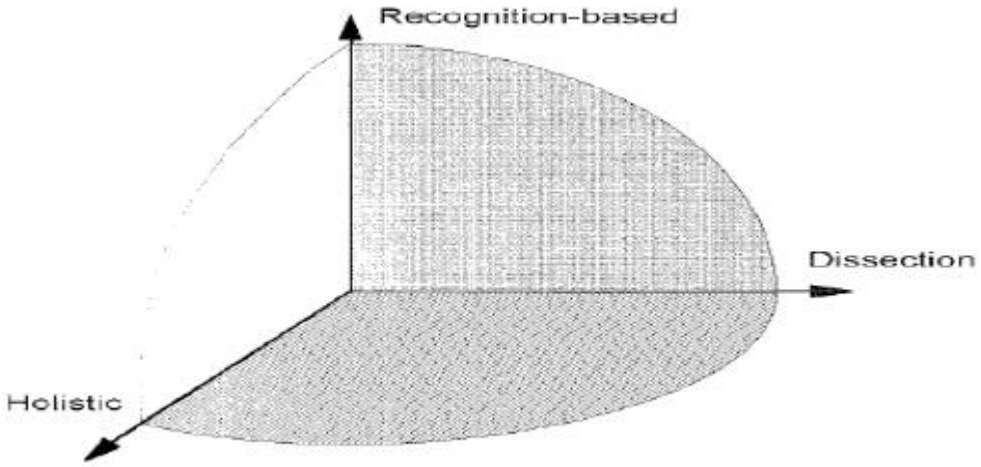


Figure 1.8: Segmentation Strategies in 3D Space [9]

The three fundamental strategies *i.e.* classical approach, Recognition based segmentation and the holistic approach can be represented in 3D space where each approach occupies orthogonal axes.

Sometimes, these strategies alone fail to separate the touching characters in 2D script. Thus, the technique can be a hybrid approach that can be the combination of either classical approach and recognition based segmentation or the combination of recognition based segmentation and holistic method and the classical approach.

The segmented image now act as input for the recognition stage which generally, assign an unknown sample to a predefined class to identify the character according to the script used, is explained in the next section.

1.8 Recognition

The OCR process assigns a character image to a class by using a classification algorithm based on the features extracted and the relationships among the features. Since members of a character class are equivalent or similar in as much as they share defining attributes, the measurement of similarity, either explicitly or implicitly, is central to any classifier.

Feature extraction is concerned with recovering the defining attributes obscured by imperfect measurements. To represent a character class, either a prototype or a set of samples must be known. The feature selection process attempts to recover the pattern attributes characteristic of each class. Global features, such as the number of holes in the character, the number of concavities in its outer contour, and the relative position of character extremities, and local features, such as the relative positions of line endings, line crossovers, and corners are commonly used. The classification stage identifies each input character image by considering the detected features.

In the statistical classification approaches, character image patterns are represented by points in a multidimensional feature space. Each component of the feature space is a measurement or feature value, which is a random variable reflecting the inherent variability within and between classes. A classifier partitions the feature space into regions associated with each class, labeling an observed pattern according to the class region into which it falls. Numerous techniques for character recognition can be investigated into four general approaches of pattern recognition, as per Arica and Vural [2], are explained here:-

1.8.1 Template Matching

The features can be as simple as the gray level image frames with individual characters or words or as complicated as graph representation of character primitives. The simplest way of character recognition is based on matching the stored prototypes against the character or word to be recognized, which template is matching. The matching operation determines the degree of similarity between two vectors in the feature space.

1.8.2 Statistical Techniques

Statistical decision theory is concerned with statistical decision functions and a set of optimality script, which maximizes the probability of the observed pattern given the model of a certain class. The measurements taken from n features of each word unit can be thought to represent an n dimensional vector space and the vector, whose coordinates correspond to the measurements taken, represents the original word unit.

1.8.3 Structural Techniques

The recursive description of a complex pattern in terms of simpler patterns based on the shape of the object was the initial idea behind the creation of structural pattern recognition. These patterns are used to describe and classify the characters in the character recognition systems. The characters are represented as the union of the structural primitives. It is assumed that the character primitives extracted from writing are quantifiable, and one can find the relations among them. The major techniques include grammatical methods and graphical methods.

1.8.4 Neural Networks

A neural network is defined as a computing architecture that consists of a massively parallel interconnection of adaptive neural processors. Because of its parallel nature, it can perform computations at a higher rate compared to the classical techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. Several approaches exist for training of neural networks.

1.9 Post Processing

The incorporation of context and shape information in all the stages of character recognition systems is necessary for meaningful improvements in recognition rates. This is done in the postprocessing stage with a feedback to the early stages of character recognition. The simplest way of incorporating the context information is the utilization of a dictionary for correcting the minor mistakes of the character recognition systems. In

post-processing, a dictionary can be used to restrict the character combinations. This can be implemented as a grammar that specifies all possible combinations of characters. The basic idea is to spell check the character recognition output and provide some alternatives for the outputs of the recognizer that do not take place in the dictionary.

Thus, the various stages of the recognition process have been discussed. Now, there arises a need to have a look on the various applications, to which the character recognition process can be applied to. The next section describes some of the applications of the character recognition process.

1.10 Applications of OCR

In recent years, OCR (Optical Character Recognition) technology has been applied throughout the entire spectrum of industries, revolutionizing the document management process. OCR has enabled scanned documents to become more than just image files, turning into fully searchable documents with text content that is recognized by computers. With the help of OCR, people no longer need to manually retype important documents when entering them into electronic databases. Instead, OCR extracts relevant information and enters it automatically. The result is accurate, efficient information processing in less time. Some of the important applications of handwritten character recognition are discussed in the following section:-

- **Banking**

  One widely known application is in banking, where OCR is used to process checks without human involvement. A check can be inserted into a machine, the writing on it is scanned instantly, and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for handwritten checks as well, though it occasionally requires manual confirmation. Overall, this reduces wait times in many banks.

- **Legal Industry**

  In the legal industry, there has also been a significant movement to digitize paper documents. In order to save space and eliminate the need to sift through boxes of paper files, documents are being scanned and entered into computer databases. OCR further simplifies the process by making documents text-searchable, so that they are

easier to locate and work with once in the database. Legal professionals now have fast, easy access to a huge library of documents in electronic format, which they can find simply by typing in a few keywords.

- **Health care**

  Healthcare has also seen an increase in the use of OCR technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. Form processing tools, powered by OCR, are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded. As a result, healthcare providers can focus on delivering the best possible service to every patient.

- **Form Processing**

  Handwritten character Recognition can be also used for form processing. Forms are normally used for collecting the public information. Replies of public information can be handwritten in the space provided and this handwritten data can be processed using the character recognition process.

- **Automated Data Entry**

  Data entry can be a time-consuming and tedious process in many offices, occupying hours of time that could be spent performing other important tasks. Data entry is essential across a wide range of industries, ranging from healthcare to finance to government agencies. Common standardized documents include insurance forms, questionnaires, bank checks, and invoices. In order to maintain well-organized and accurate records, most businesses and organizations must update their databases regularly with information from incoming forms and documents.

- **OCR in other fields**

  OCR is widely used in many other fields, including education, finance, and government agencies. OCR has made countless texts available online, saving money for students and allowing knowledge to be shared. Invoice imaging applications are used in many businesses to keep track of financial records and prevent a backlog of

payments from piling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes. As the technology continues to develop, more and more applications are found for OCR technology, including increased use of handwriting recognition. Furthermore, other technologies related to OCR, such as barcode recognition, are used daily in retail and other industries.

1.11 Terminology Used

- *Binarization:-* It is the process of conversion of the grayscale or color images into binary images by picking a threshold value, which reduces the storage requirements.
- *OCR:-* It is the process of transforming the graphical marks associated with human handwritten document into the symbols that are stored on a computer system in the form of 8-bit ASCII code or 16-bit Unicode.
- *Segmentation:-* It is a process that seeks to decompose an image of a sequence of characters into sub images of individual symbols by segmenting lines and words. In other words, it decomposes a document into its sub components.
- *Water Reservoir*:- The water reservoir can be defined as if water is poured from one side of a component, the cavity regions of the component where water will be stored are considered as reservoirs. There can be top, bottom, left, and right reservoirs based on the side of the cavity.
- *Recognition*:- Recognition is defined as to extract the information from the segmented character.

This concludes the introduction to the character recognition process. To understand the current state of art, the literature has been reviewed and presented in the next chapter.

CHAPTER 2

LITERATURE SURVEY

The inexplicable research work in character recognition area has led to the development of innumerable approaches to deal with the various aspects of the character recognition. The approach for the character recognition process may vary according to the script under consideration. In India, many official languages such as Hindi, Marathi, Sindhi, Nepali and Sanskrit are from the Devnagari script. Across central and northern parts of India, more than 300 million people use Devnagari script for documentation. In order to understand the current state of art in this area, a survey of work done related to the character recognition process has been presented in this chapter.

Lu [1995] has stated overview of the character segmentation techniques in machine-printed documents. So far, in most Optical Character Recognition (OCR) systems, either commercial products or systems described in the published literature, Recognition algorithms are developed on isolated characters. Character segmentation is all too often ignored in the research community, yet broken and touching characters are responsible for the majority of errors in the automatic reading of both machine-printed and handwritten text. It covers techniques for segmenting uniformed or proportional fonts, broken and touching characters; techniques based on text image features and techniques based on recognition results.

Casey and Lecolinet [1996] provided a review of various techniques and methodologies in character segmentation. The importance of segmentation in the recognition process and various steps of classical optical character recognition process have been discussed. They divided the segmentation strategies into classical approach, recognition based approach and holistic approach. The classical approach that identifies the segments based on character like properties has been discussed. The recognition based approach that finds those components in image that matches the classes in alphabet and the holistic method recognizes the word as a whole has also been explained in the paper. The dissection

techniques like projection analysis, white space and pitch approach and connected component processing have been discussed.

Trier *et al.* [1996] presented an overview of feature extraction methods for the offline recognition of the isolated characters. Different feature extraction methods, which are designed, for the different representation of the characters, such as solid binary characters, character contours, skeletons, gray scale images of each character have been discussed. The feature extraction methods have been explained in terms of invariance properties, expected distortion, re-constructability and variability of the characters. The various processing steps of the optical character recognition process, namely gray level scanning, binarization, segmentation, representation, feature extraction, recognition and contextual processing have also been stated.

Ha and Bunke [1997] presented a new approach to offline handwritten numeral recognition. They proposed a recognition method, which is able to account for a variety of distortions due to eccentric handwriting. The technique for the perturbation based recognition system has also been discussed. The key idea of the perturbation approach, which lies in the process of reversing an input image back to one of its standard forms, has also been stated. The methodology for the parameterization process based on four geometric transformations, namely, rotation, slant, perspective view and shrink has also been explained. The approach to replace normalization by a set of perturbation processes modeling writing habits and instruments has also been discussed.

Mo and Mathews [1998] proposed an adaptive filter to be used prior to binarization for the edge enhancement of the characters. The importance of edge enhancement and noise reduction in the document image for the recognition process has also been explained. The paper also stated the similarity of proposed approach with the equalization of the binary communication channels. An introduction to the quadratic filter for the removal of the noise from the document acquired during image acquisition process has also been given. The mathematical description for the quadratic filter and its application has also been given. The design and implementation issues with the proposed algorithm have also been

stated in the paper. The low pass and high pass filters and their application for binarization process has also been summarized.

Cai and Liu [1999] proposed an approach that integrates the statistical and structural information for unconstrained handwritten numeral recognition. The approach that uses the state duration adapted transition probability to improve the modeling of state duration in conventional markov models and uses macro states to overcome the difficulty in modeling pattern structures by markov models has also been explained. The technique for the encoding of the orientations into discrete codebooks and the distributions of locations are modeled by joint Gaussian distribution functions has also been discussed. The preprocessing methods for the disjoint connection region, slant correction, size normalization have also been dealt with.

Lehal and Singh [2000] proposed a system for the recognition of the machine printed characters of the Gurmukhi script. The various characteristics of Gurmukhi script have been stated. The preprocessing techniques like skew correction and detection, thinning and smoothing the headlines have been discussed. The horizontal and vertical projection profiles of the characters and their usage for line segmentation and word segmentation has been explained. The character segmentation based on connected component identification has also been discussed. The recognition process based on feature extraction and classification of sub symbol has also been described. The set of simple features used for binary decision trees and nearest neighbor classification techniques used for classification and recognition, has also been explained.

Arica [2001] has concluded that Character recognition (OCR) has been extensively studied in the last half century and progressed to a level sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present OCR methodologies and creates an increasing demand on many emerging application domains, which require more advanced methodologies. This material serves as a guide and update for readers working in the OCR area. First, the historical evolution of OCR systems is presented. Then, the available OCR techniques

with their superiorities and weaknesses are reviewed. Finally, the current status of OCR is discussed, and directions for future research are suggested. Special attention is given to the off-line handwriting recognition since this area requires more research to reach the ultimate goal of machine simulation of human reading.

Kasturi *et al.* [2002] gave a detailed description of the document image analysis process. The sequence of steps starting from data capture, pixel level processing, feature level analysis until text recognition and analysis has been elaborated. A brief analysis of graphical documents has been presented. The techniques for noise reduction and binarization have been discussed. The techniques for thinning and region detection, chain coding and vectorization have also been explained. The techniques for line and curve fitting, OCRitical point detection, skew estimation, layout analysis have also been discussed. The strategy for feature extraction and classification based on template matching and contextual processing has been explained. The various OCR's for Indian languages and document analysis in multilingual context has also been stated.

M. Blumenstein [2002] has concluded that new preprocessing techniques used in a system for offline handwritten word recognition have been presented. Techniques for underline removal and skew detection have been proposed and tested on a benchmark database of handwritten words. Underline removal was successful in 97.16% of cases and word skew was correctly determined in 96.12% of cases. In future, the baseline estimation algorithm shall be updated to take into account those words that contain abnormal horizontal strokes. An extra step shall also be included in our word recognition system to remove excessive noise that adversely affects our algorithms.

Pal *et al.* [2003] presented an automatic scheme to identify text lines of different Indian scripts from a document. The method for grouping the scripts into some classes in accordance with the script characteristics has also been stated. The various features used for script identification, namely headline feature, horizontal projection profile, water reservoir principle based feature, left and right profile, feature based on jump continuity have also been explained. The technique for the identification of the script based on the

various features has also been discussed. The approach is insensitive to font, style and case variation has been stated.

Blumenstein *et al.* [2004] presented a technique for cursive character recognition applicable to segmentation based word recognition systems. The proposed novel feature extraction technique that extracts direction information from the structure of character contours has also been discussed. The methodology for the extension of the principal so that the direction information has integrated with a technique for detecting transitions between background and foreground pixels in the character image has also been explained. The technique to find the direction feature through determination of directions and formation of feature vector has also been stated. The proposed technique has also been compared with the standard direction feature extraction technique and an improvisation has been found.

Jayarathna and Bandara [2006] presented an approach for the segmentation of offline handwritten connected two digit strings. They developed an algorithm, which provides a solution based on the analysis of the foreground pixel distribution to segment the connected digit string pairs. The junction based splitting technique, which decides complete segments of the connected digit strings, has also been explained. The use of the fuzzy characteristic values at the merging of the complete segments that isolates the major segments from the minor segments has also been discussed. The process for binarization, isolation of connected character skeleton into correlation area, junction based segmentation, identification of starter points and junction points, traversal through the correlation area and merging of segments aOCRoss junctions has also been stated.

Arora *et al.* [2007] presented a two stage classification approach for handwritten Devnagari character. The first stage that uses the structural properties like shirorekha, spine in character has been discussed .The second stage which exploits some intersection features of characters which are fed to a neural network has also been explained. The preprocessing stage including size determination, distortion removal and normalization has also been explained. They stated that the simple histogram based method does not

work for finding shirorekha, vertical bar in handwritten Devnagari characters. They designed a differential distance based technique to find a near straight line for shirorekha and spine in the characters.

Banashree and Vasanta [2007] proposed a recognition scheme for handwritten Hindi numerals. The work focused on a technique in feature extraction i.e. global based approach using end points information, which is extracted from images of isolated numerals has been explained. These feature vectors are fed to neuromemetic model that has been trained to recognize a Hindi numeral. In the proposed scheme, data sets are fed to neuromemetic algorithm, which identifies the rule with highest fitness value and the template associates with this rule has nothing but identified numerals. A global based approach using end points information for feature extraction has been used.

Loh Zhi Chang [2008] stated that signs are everywhere in our daily life. They can pose a danger if people do not understand them. Thus, it will be beneficial to have an OCR application on a mobile phone. OCR application on mobile devices is no longer a dream due to the advancement in mobile technology. There are many various studies about embedded Optical Character recognizer. Although there are many OCR research ongoing, most of the research is focus on the OCR engine of the application. Pre-processing stage is often overlooked by many research works. Pre-processing step optimizes the images. Before the actual processing such as text detection and recognition is being carried out. In this work, two preprocessing steps are implemented, scanning algorithm and Otsu's threshold. Scanning algorithm helps users to define the ROI of the image while the Otsu's threshold helps to differentiate between the background and the foreground. Experiments are carried out with the OCR application being tested with 40 sign images. From the experiments, it is shown that the pre-processing methods do improved the accuracy of an OCR engine overall. By defining the ROI of an image, noises in the background will be eliminated. This increases the accuracy of the OCR engine. This method can be done by requesting user input and scan the nearby region for the edges of the sign. The result also shows that Otsu's method of thresholding helps

convert the grayscale image to binary image. This results in a better recognition of text as the noises in the background will be minimized or even eliminated.

Pal *et al.* [2008] proposed a quadratic classifier based scheme for the recognition of offline handwritten characters of three popular south Indian scripts namely, Kannada, Telugu, and Tamil. The technique to obtain the features used from the directional information has been stated. The method for feature computation that includes the segmentation of the bounding box of a character into blocks and the computation of the directional features in each block has also been discussed. The methodology for down sampling the blocks by a Gaussian filter and feeding the generated features from the down sampled blocks to a modified quadratic classifier for recognition has also been explained. They have used two sets of features namely, 64 dimensional features for high speed recognition and 400 dimensional features for high accuracy recognition.

Ramteke and Mehrotra [2008] employed a method based on invariant moments and the divisions of numeral image for the recognition of handwritten Devnagari numerals. The technique has independent of size, slant, orientation, translation and other variations in handwritten characters. The technique for normalization to a fixed pixel size for each individual image after the segmentation has also been explained. Seven central invariant moments have been evaluated for each character and its parts by dividing it by three different ways. In all, there are 78 features corresponding to each character. The Gaussian distribution function has been adopted for classification. The method to separate a text line from the previous and following lines by white space with horizontal projection has been stated.

Syed *et al.* [2009] have explored the sensitivity of fixing free parameters values for all pixels of a camera-captured document image and demonstrated that no matter how to find the free parameters values (either manually or automatically), some range of values of free parameters gives better binarization results for foreground (text-area) document image regions and some other range of values gives better binarization result for background regions. Also overcome this sensitivity by introducing the idea of not using the constant values of free parameters for all pixels, but use different values of free

parameters for pixels belong to roughly estimated foreground and background regions. For this purpose they presented the idea of using multi-oriented multi-scale anisotropic Gaussian smoothing and ridges detection for roughly estimating foreground regions from grayscale document image. Execution time of our method is quite slow as compared to other locally adaptive binarization methods because of the approximation of foreground regions before applying local binarization method.

Yasser *et al.* [2009] concluded that preprocessing techniques used in document images as an initial step in character recognition systems were presented. Future research aims at new applications such as online character recognition used in mobile devices, extraction of text from video images, extraction of information from security documents and processing of historical documents. Even though many methods and techniques have been developed for preprocessing there are still problems that are not solved completely and more investigations need to be carried out in order to provide solutions. Most of preprocessing techniques are application-specific and not all preprocessing techniques have to be applied to all applications. Each application may require different preprocessing techniques depending on the different factors that may affect the quality of its images, such as those introduced during the image acquisition stage. Image manipulation/enhancement techniques do not need to be performed on an entire image since not all parts of an image is affected by noise or contrast variations; therefore, enhancement of a portion of the original image maybe more useful in many situations. This is obvious when an image contains different objects which may differ in their brightness or darkness from the other parts of the image; thereby, when portions of an image can be selected, either manually or automatically according to their brightness such processing can be used to bring out local detail. In conclusion preprocessing is considered a OCRucial stage in most automatic document image analysis systems and without it the success of such systems is not guaranteed.

Chang, *et al*. [2010] has concluded that Document images with non-uniform brightness require binarization methods with delicate local thresholds that must be determined according to various conditions. For this purpose, we propose a region-based binarization method. Binarization based on information provided by a region is effective and robust,

provided that the SVM method is used to construct decision functions from the information provided by training samples. The experiments produce favorable results with respect to the thresholding actions judged in terms of the visual quality of images, and also in terms of the OCR performance.

Singh *et al.* [2010] has proposed that there are about 300 million people in India who speak Hindi and write Devnagari script. Research in Optical Character Recognition (OCR) is popular for its application potential in banks, post offices, defence organizations and library automation etc. However most of the OCR systems are available for European texts. They proposed a technique for OCR System for different five fonts and sizes of printed Devnagari script using Artificial Neural Network. The recognition rate of the proposed OCR system with the image document of Devnagari Script has been found to be quite high

Patnaik *et al.* [2011] has given a technique for the evaluation of binarization algorithms. This technique is appropriate for document images that are difficult to be evaluated by techniques based on only segmentation or recognition of the text. In order to demonstrate the proposed method, tested three existing binarization algorithms, performed experiments on 100 document images. Although there is a better performance of the Adaptive binarization algorithms compared to other, the other ones have produced almost similar results.

Shukla and Banka [2011] concluded a segmentation scheme for the recognition of the printed Devnagari script. The features of the Devnagari script such as basic characters, modified characters have been explained in the paper. The preprocessing steps consisting of binarization, noise removal, skew detection and correction have been discussed. The segmentation process consisting of the line segmentation, word segmentation and character segmentation has also been elaborated upon. The authors have separated the individual line in a script document image based on the peak of the horizontal histogram. The technique used for the word segmentation works on the basis of the vertical histogram of the extracted line.

Ramamurthy *et al.* [2012] presented the issues in recognizing the Devnagari characters in the wild like sign boards, advertisements, logos, shop names, notices, address posts. A detailed study of the Devnagari character recognition using the state of art character recognition and object recognition tools has been carried out by them. A description of the Devnagari script consisting of consonants, vowels, conjuncts has been given. The preprocessing techniques smoothing and binarization have been discussed. The features like pixel density, directional features, histogram of oriented gradients, shape context have been explained by the authors. The performance evaluation of the existing state of art features and classifiers on specific database is presented.

Cuc *et al.* [2012] presented the handwritten digit recognition on well known image databases using state of the art feature extraction and classification techniques. The preprocessing technique that includes conversion into gray scale image, conversion into binary image, morphology method and normalization has been explained. The feature extraction methods namely, seven moments and image averaging has also been discussed. The classification methods including neural network and template matching has also been elaborated upon.

The literature survey for the optical character recognition process has been presented in this chapter. Based on this survey, the problem definition and justification the have been discussed in the next chapter.

CHAPTER 3

PROBLEM STATEMENT

3.1 Problem Definition

The optical character recognition process translates the scanned image into a machine editable form or electronic format. This process has different phases which already have been discussed in chapter 1. This process can be used to translate books, articles and documents etc. into editable form. In OCR process, first step is scanning that is used to convert a paper document into an image document. The image resulting from the scanning process may contain a certain amount of noise. It may be that the Scanned Document may have varying intensity or skewed. That may affect the OCR system and may cause not accurate recognition. It is quite evident from the literature survey that numerous approaches have been developed for the handwritten numeral recognition process. However, it can also be observed that the handwritten numeral recognition is still a fascinating area for the researchers to design a robust and efficient algorithm for the same. The numerous techniques that have been developed are able to provide good recognition rate ranging from 85% to 90% but these techniques have some implementation gaps. Most of the approaches for recognition process are based on the neural network technique and its adaptations. These techniques are computationally difficult and require good amount of time to perform the training of the systems in order to provide good results.

3.2 Justification

Depending on the resolution on the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a preprocessor to smooth the digitized characters. Based on these OCR stages, the proposed work has been discussed in three sections as preprocessing in 4.1, segmentation in 4.2, size Normalization in 4.3 and recognition in 4.4.

In preprocessing phase, modified binarization algorithm has been proposed to convert a gray scale image into a binary image. This Modified binarization algorithm will give a accurate result on those images also in which pixel values varying a lot. After modified binarization algorithm author has proposed a algorithm for the skew correction. Preprocessing is a very first and very essential part of OCR. The output of this phase will be the input of segmentation phase. If preprocessing has not been done correctly the right segmentation and recognition is not possible. It can be observed that many numeral writings are present in these applications and the recognition of the numerals will ease out the working of these applications. Thus, the recognition of the numerals is a significant area to work on. Moreover, the numeral recognition can be integrated with several other fields to help with several applications like reading details like postal zip code, employee code, passport number and processing of forms and bank cheques. It can also be utilized in schools and colleges to process the list of marks. Thus, there is a need for a simpler approach for the character recognition process. The authors have thus, tried to develop a simple and efficient algorithm for the preprocessing, segmentation, size normalization and English numeral recognition, which has been discussed in the nest chapter.

CHAPTER 4

PROPOSED SOLUTION

The problem which has been discussed in the previous section, a number of techniques has been developed to give the better result and to improve the performance of OCR. The handwritten numeral recognition process includes a number of stages, as discussed in chapter 1. Based on these stages, there may be some difficulties as discussed in section 3.1.Based on the problem author has proposed some algorithm. The proposed work has been discussed in three sections as preprocessing in 4.1, segmentation in 34.2, Size Normalization in 4.3 and recognition in 4.4. The section related to preprocessing deals with the cleaning up of the image by means of modified binarization algorithm and skew correction techniques. The section pertaining to segmentation discusses the segmentation of the image into sub components by using line segmentation, word segmentation and character segmentation. After the character segmentation, size normalization technique is used to increase or decrease the size of individual character. The section dealing with recognition describes the classification of the image based on the features extracted from the image. The features used are water reservoirs, closed loops and sidebars. The classification structure based on these features is used to perform the recognition.

4.1 Preprocessing Techniques

Preprocessing aims to produce the data so that data is easy for the system and can operate accurately. It includes technique to remove the noise and give a noise free data for the next phase segmentation. The preprocessing technique are improved average binarization algorithm and skew correction.

4.1.1 Binarization Algorithm

A **binary image** is a digital image that has only two possible values for each pixel. Typically the two colors used for a binary image are black and white. binarization is used as a pre-processor before Character Recognition. Image binarization converts an image (up to 256 gray levels) to a black and white image (0 or 1).

Approach Used:-

Here author has used two algorithms for the binarization. First one is Modified Average Binarization algorithm and the second one is Adaptive Binarization algorithm. Adaptive Binarization algorithm is the modified form of Modified Average Binarization method, which gives a better performance

- Modified Average Binarization method
- Adaptive Modified Average Binarization method

Modified Average Binarization Algorithm

Modified Average Binarization method extends average binarization method to a novel modified binarization scheme. The steps for the algorithm are defined as follows:

1: Calculate the intensity of every pixel.

2: Calculate the total sum of intensity of every pixel.

3: Calculate the total number of pixel in the image.

4: Find the average intensity value *K*.

5: Find the number of highest and number of lowest intensity value.

6: divide the number of highest and number of lowest and set it to *t*.

7: Multiply *t* and *k* and store into *k.*

8: if intensity of the pixel is less than *k* then set it to black else white.

9: Repeat the all steps till the end of the image.

10: Exit.

Adaptive Modified Average Binarization Algorithm

Adaptive average binarization method extends average modified binarization method to a novel adaptive binarization scheme. The first step of our method is to divide images into NxN blocks, and then average binarization method is applied straightaway in each of the blocks. Then each and every pixel is applied with a nonlinear quadratic filter to fine tune all the pixels according to the local information available. Adaptive Binarization technique combines global thresholding using local information to fine tune the pixel.

The steps for Adaptive Average Modified Average Binarization method are as follows:

1: Divide images into N x N blocks.

2: Apply Average Binarization method for each block.

3: Combine all the local information to find the global information.

4: Repeat the all steps till the end of the image.

5: Exit.

4.1.2 Skew Correction

During the procedure of scanning input, for some unavoidable reasons, images would bring in Skew issue. No matter scripts is very sensitive to the Skew of images, which is quite important to the document skew detection of Preprocessing.

In the given method, the image will be rotated by α degree through coordinate transformation by using the given formula.

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = TX_0 = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

Skew Correction Algorithm

1: Find the coordinates value of every pixel.

2: Draw a virtual line A as explained in next section segmentation.

3: Draw another straight line B.

4: Find the angle between Line A and Line B and set angle into α.

5: Apply the above formula to get the new coordinate value.

6: Exit.

4.2 Segmentation Technique

Segmentation basically deals with the decomposition of the document into sub components like lines, words and characters. The extent of accuracy of segmentation directly impacts the accuracy of recognition; hence, segmentation is very important stage of the handwritten numeral recognition process. The algorithm to perform segmentation is described in this section.

Algorithm for Segmentation

1: Calculate the height and width of the image and store it into *h* and *w***.**

2: Search a BLACK pixel from 0 to *w in* X coordinate**.**

3: If a pixel is BLACK then *h* will be set as $h=h+1$ and repeat step 2.

4: Else draw a line and search for the BLACK pixel.

5: Repeat Step 3 for all the pixels.

6: Draw a box for every Line.

7: Search a BLACK pixel from 0 to height of the window in Y coordinate**..**

8: If a pixel is BLACK then *x* will be set as $x=x+1$ and repeat step 7.

9: Else draw a line and search for the BLACK pixel.

10: Repeat Step 7 for all the pixels.

11: Exit.

4.3 Size Normalization

The result from the character segmentation stage provides isolated characters which are ready to be passed into the recognition stage, therefore, the isolated characters are normalized into a specific size, decided empirically or experimentally depending on the application and the feature extraction or classification techniques used, then features are extracted from all characters with the same size in order to provide data uniformity. Sometimes we need to normalize the data only horizontally or only vertically according to the size of character. So size normalization can be done in 3 ways:-

- Normalization
- Horizontal Normalization
- Vertical Normalization

Algorithm for Normalization:

1: Find the height and width of the image.

2: To increase the height and width by multiply it by 1.2.

3: To reduce the height and width by divide it by 1.2.

Algorithm for Horizontal Normalization:

1: Find the height and width of the image.

2: To increase width multiplies it by 1.2.

3: To reduce width divides it by 1.2.

Algorithm for Vertical Normalization:

1: Find the height and width of the image.

2: To increase height multiplies it by 1.2.

3: To reduce height divides it by 1.2.

4.4 Recognition Technique

The segmented characters obtained from the application of the segmentation algorithm are now required to be recognized and thus, acts as input for the recognition stage, which is discussed in the following section.

Recognition refers to the process of finding out the predefined class to which the unknown sample belongs, in accordance with the script used. It basically consists of two phases, feature extraction and classification.

Firstly, the required features are extracted from the images and then, based on those features, a classification structure is developed. The classification structure defines an appropriate class for each image based on the features present in the image. The features used in the current algorithm are presence of sidebars, presence of closed loops and presence of water reservoirs in the characters.

The present algorithm for recognition derives its basis from the concept explained in Kaur *et al.* [16]. They have used the water reservoir analogy principle and various features to form a classification structure, which assigns the extracted fragment to a particular class.

An effort has been to apply the same principle for the recognition stage. The water reservoir can be defined as if water is poured from one side of a component, the cavity regions of the component where water will be stored are considered as reservoirs.

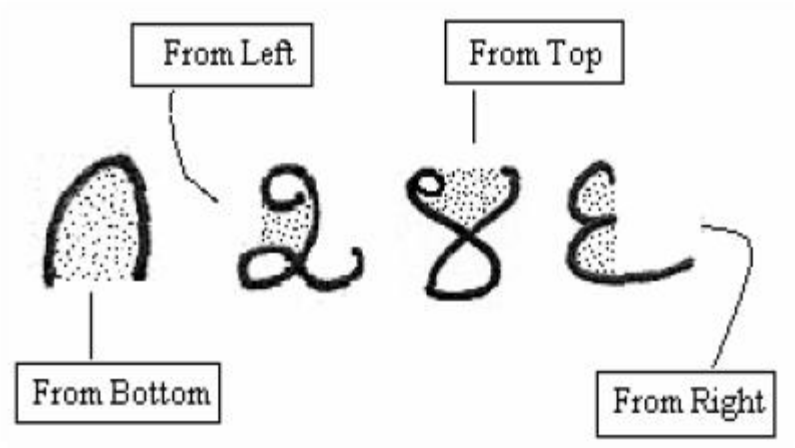


Figure 4.1: Water Reservoirs in Numerals

The left (right) reservoir is the water stored cavity regions of the component, when water is poured from left (right) side of the component. Similarly, top and bottom reservoirs can be defined. The top, bottom, left, and right reservoir of numerals are illustrated in figure 4.1.

The classification structure used in the algorithm is shown in figure 4.2. The recognition algorithm based on these features and this classification structure, is presented in this section. In figure 4.2, there is a flow chart to describe that how recognition has been done in a simple and efficient manner. It has been implemented for the English Handwritten numeral.

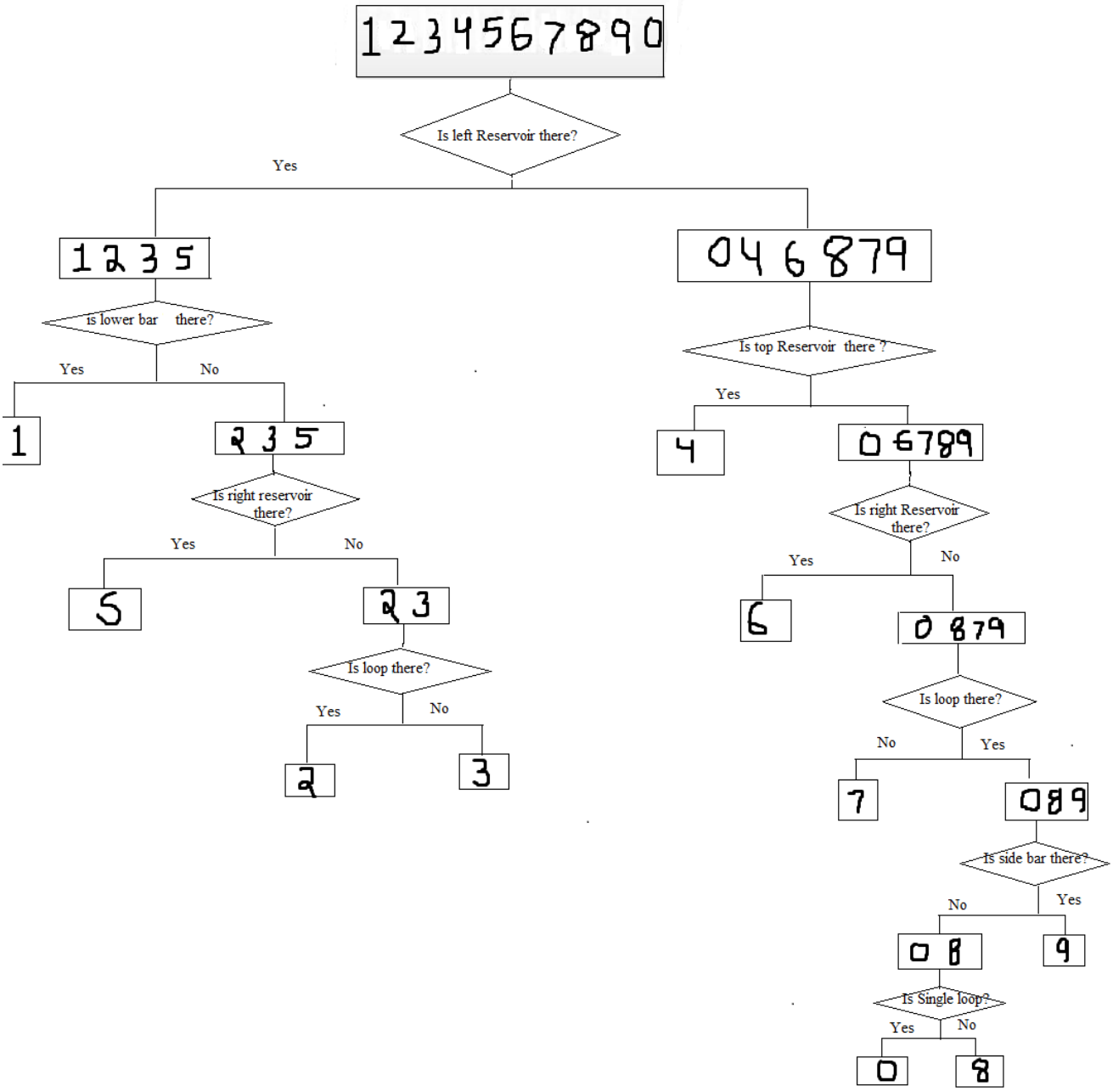


Figure 4.2: Classification Structure of English Numerals

The proposed algorithms have been implemented and the results are discussed in the next chapter.

CHAPTER 5

RESULTS AND DISCUSSION

The proposed work discussed in chapter 4 has been implemented by the authors in the JAVA language using eclipse platform. The algorithms proposed are implemented and tested on several documents. Initial step is preprocessing then output of preprocessing is the input of next Segmentation Phase. Then Segmented Character is the input of the recognition phase to recognize the characters. The current section describes the outputs obtained for the stepwise execution of the algorithms. The results obtained are also discussed in this section.

Results

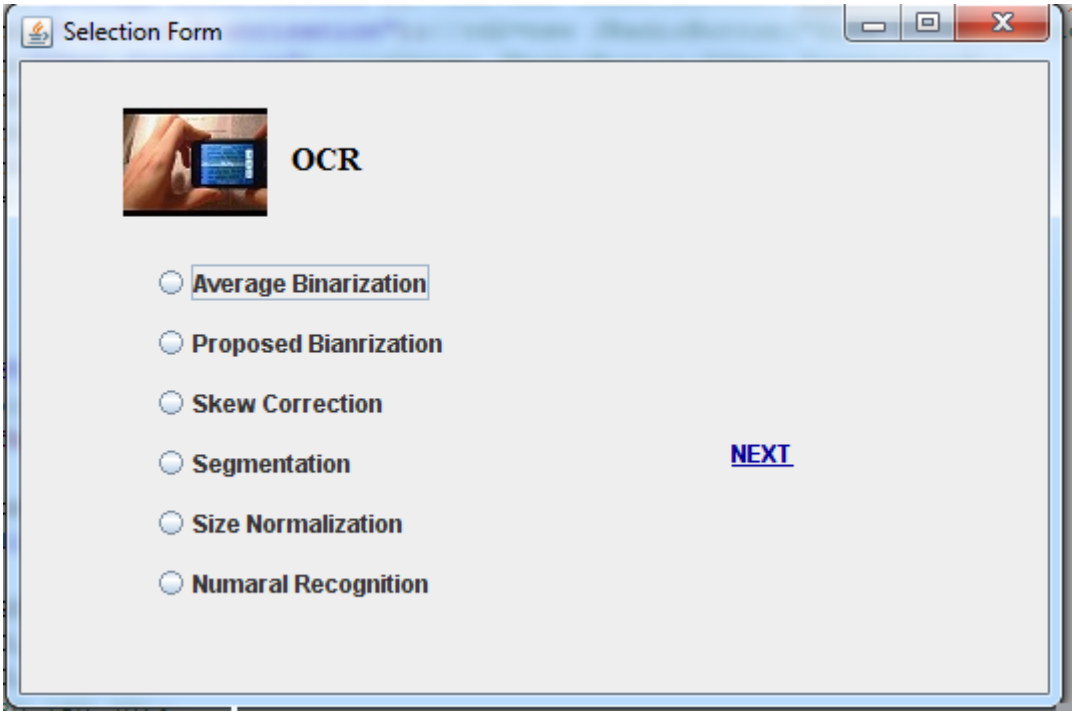


Figure 5.1: Selection Form

The initial step is to select Average binarization radio button from selection form (Figure 5.1) having different Preprocessing Techniques. Any one of them can be selected. Average Binarization has been selected then clicked on the Next link to proceed further process.

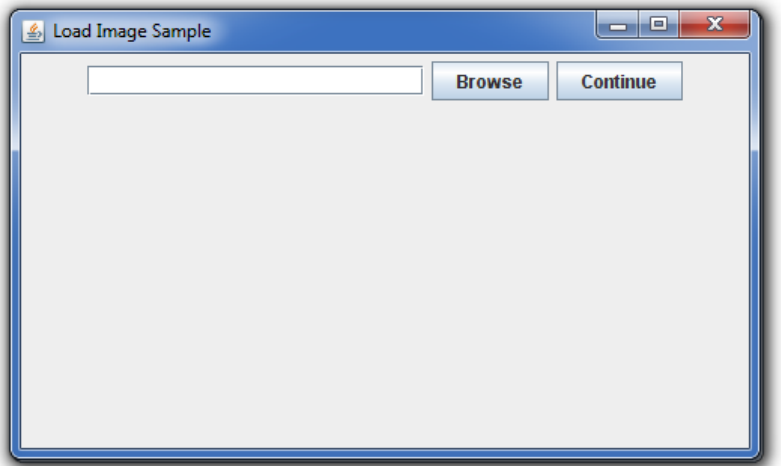


Figure 5.2: Upload Image

A new window is opened as shown in figure 5.2. Here are two buttons Browse and Continue. On clicking Browse button any image, can be uploaded.

Figure 5.3: Uploading Image

Then, on choosing any image and clicking on Open button, any image can be uploaded. It can be canceled also to click on the Cancel button as shown in figure 5.3. The image has been uploaded successfully. Here the path of the image has been printed in the Label. Now next step is to click Continue button for the further process, to apply Average Binarization, which is a Preprocessing Technique.

5.1 Average Binarization Result

Uploaded Gray scale image has been converted into binary image.

Example 1

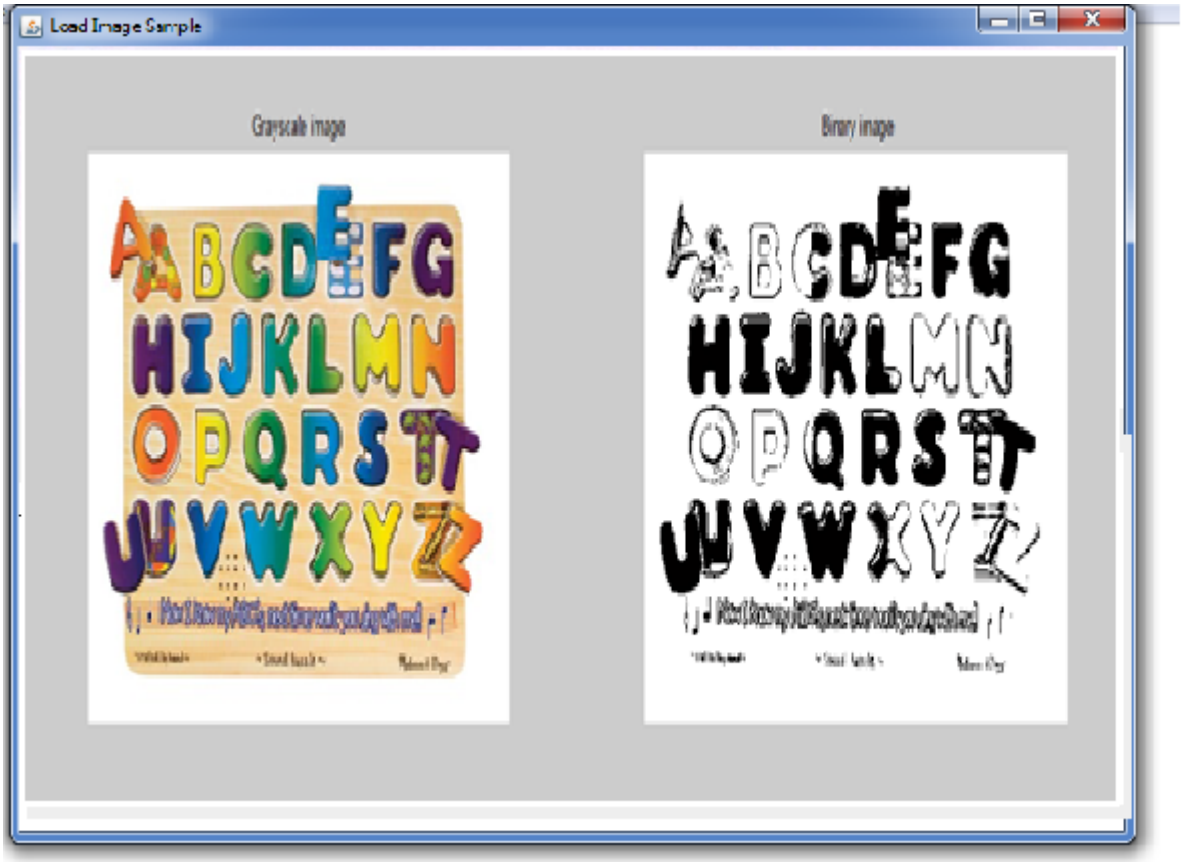


Figure 5.4: Binarization using Average binarization Method

The above Left side image is the original image and the right side image is the image on which average binarization technique has been applied. The algorithm as explained in

section 4.1 will convert a color image into a binary image but some captured or scanned images having a lot of variation in intensity value. In that case, when we apply this average binarization method, it will not give the accurate result.

Average Intensity depends upon the intensity value of the all the pixels of the images. Some is very high and some are average. Average binarization method will not give the satisfactory result and it may affect the OCR system to give the accurate result.

Example 2

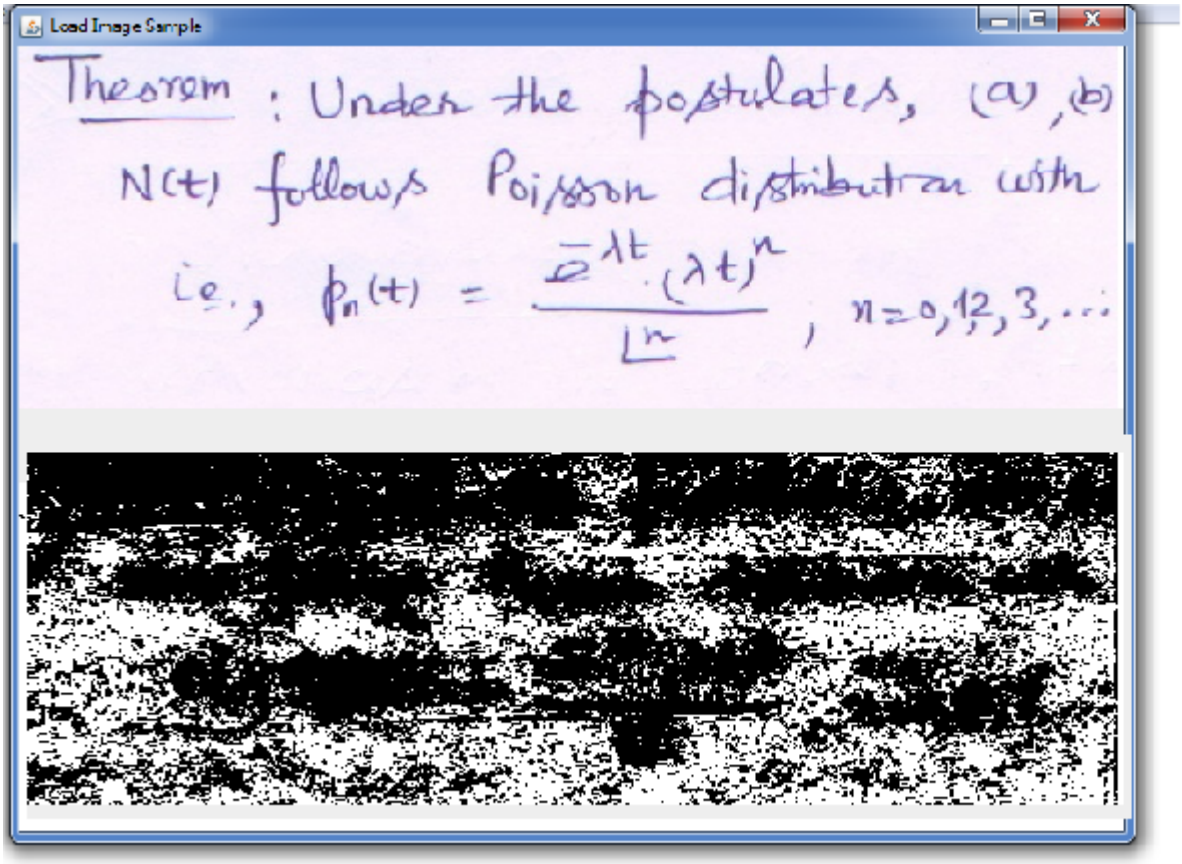


Figure 5.5 Binarization using Average binarization Method

The above image is a scanned image of a hand written document in which intensity of each pixel is varying a lot from each other. Some pixel values are very high, Some pixel values are very low. Average value is very low because the pixels having the very low intensity value are 100 times more than the pixel having high intensity value. So the

average value of all the pixels is very low and almost all pixels color has been changed into BLACK.

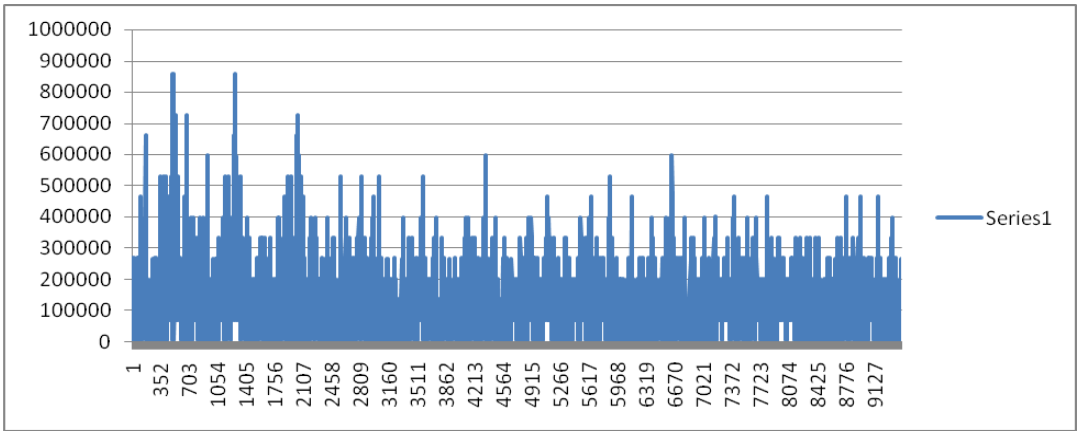


Figure 5.6: Variation among Pixels of image 5.5

We can see in the above graph of the image 5.5. Figure 5.5 is a image of a camare captured image in which among the all pixels, there is a lot of variation. When author apply Average Binarization method, it will not give the satisfactory resul as we can see in the Figure 5.5.

5.2 Proposed Average Binarization Algorithm result

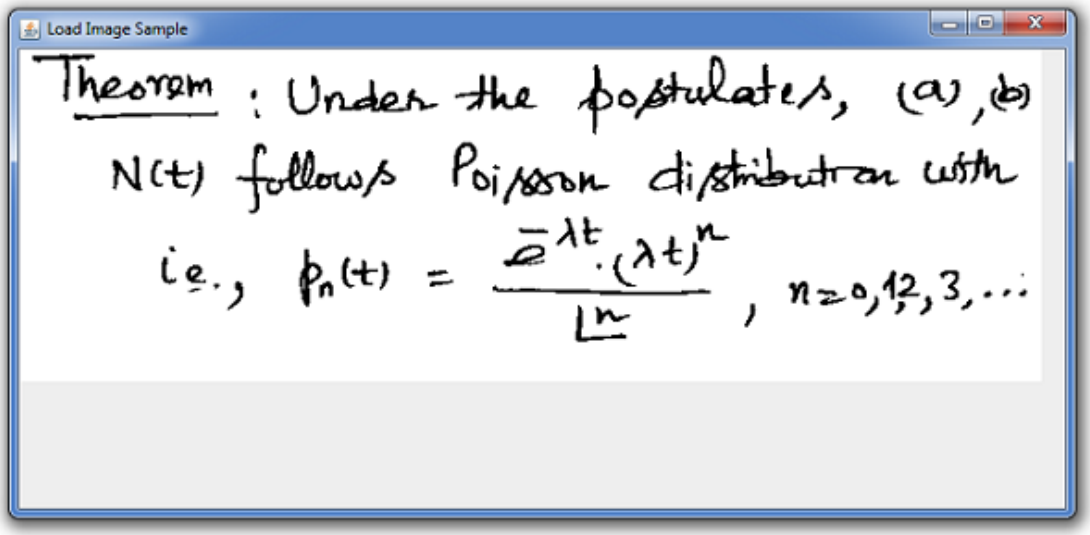


Figure 5.7: Proposed Average Binarization Result

The proposed averaged binarization algorithm which is described in the above 4.2 section is tested on text images in comparison with average binarization algorithm. The original image is shown in Figure 5.4. Result of Average binarization algorithm output image is shown in Figure 5.7. In proposed algorithm, average value has been changed accordingly to the scanned or camera captured image. Now not almost every pixel is black because average value is increased according to the image and it has given the more accurate result.

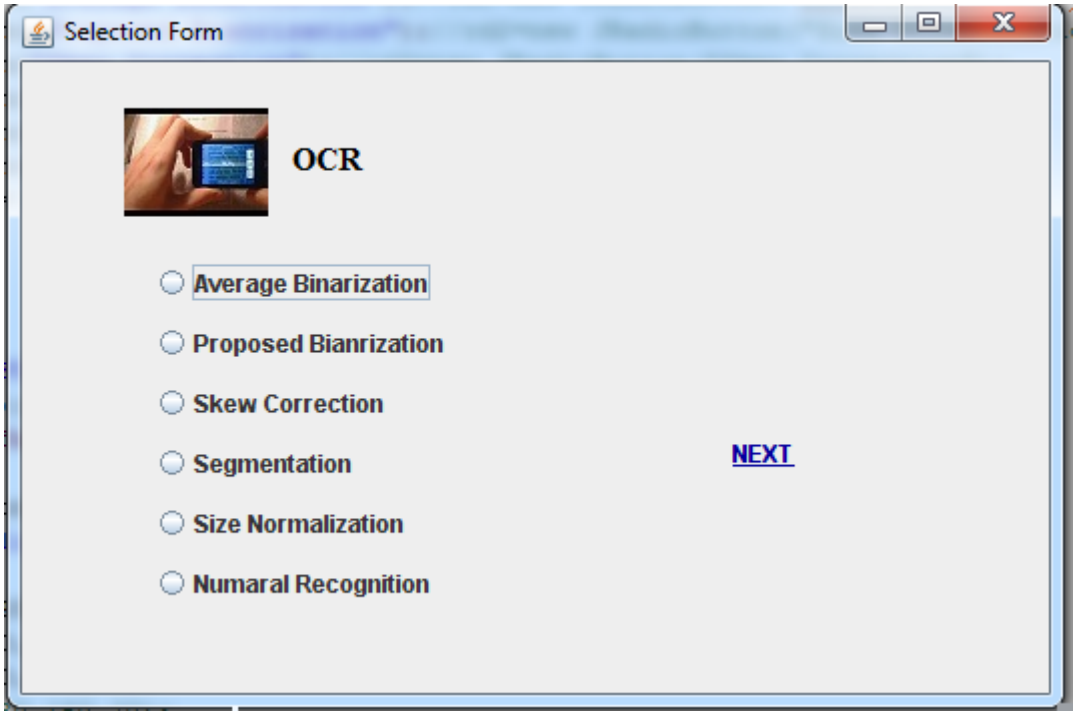


Figure 5.8: Selection Form

5.3 Skew Correction Result

Come back on this Selection Frame and Select the Skew Correction technique and click on NEXT link. A new window will be opened. Here are two buttons again Browse and Continue as we can see in Figure 5.2. On clicking Browse button can be uploaded any image. Then choose any image and click on Open button to upload the image. It can be canceled also to click on the Cancel button as in Figure 5.3.

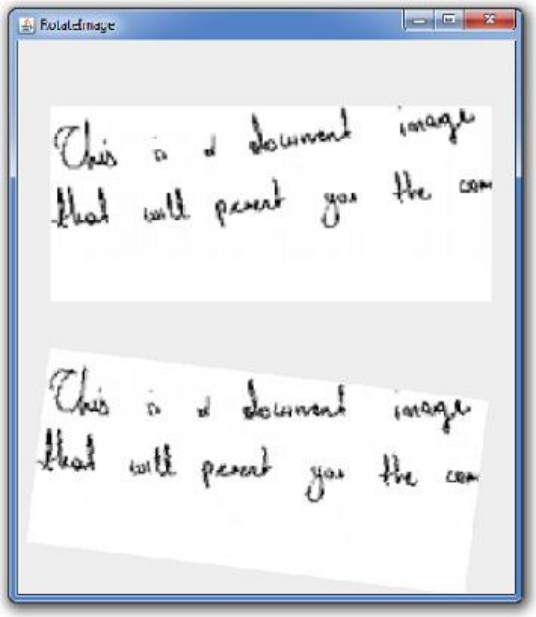


Figure 5.9 Skew Normalization

The above image is before the skew correction and below image is after skew correction. The Skew Correction algorithm has been defined in the above section 4.1, which has been implemented and giving the result which is the input for the next phase that is a Segmentation Phase.

5.4 Segmentation Phase Result

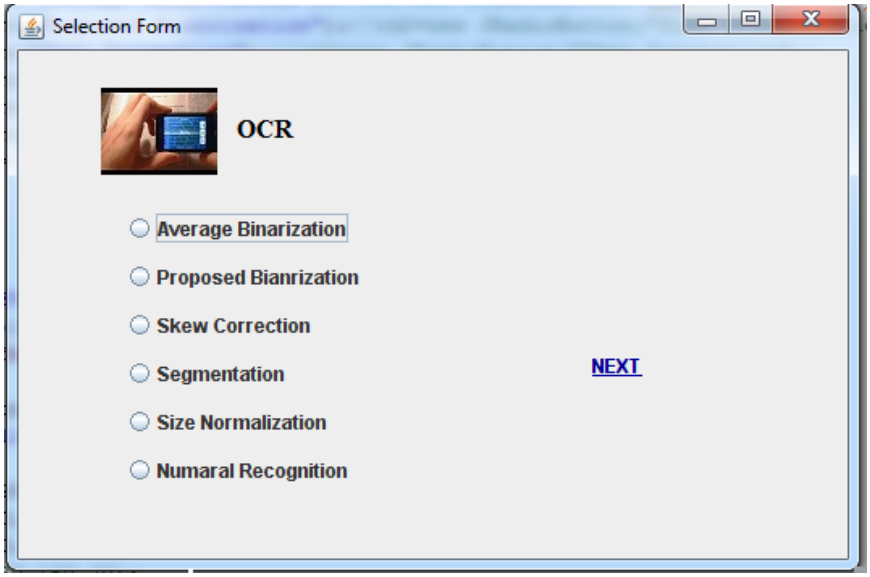


Figure 5.10: Selection Form

Now Form the selection Form as shown in figure 5.10, After Preprocessing Segmentation Phase has been applied. The Output of the preprocessing phase is the input of segmentation phase. Here selecting on Segmentation radio button n click on the NEXT link, segmentation has been done.

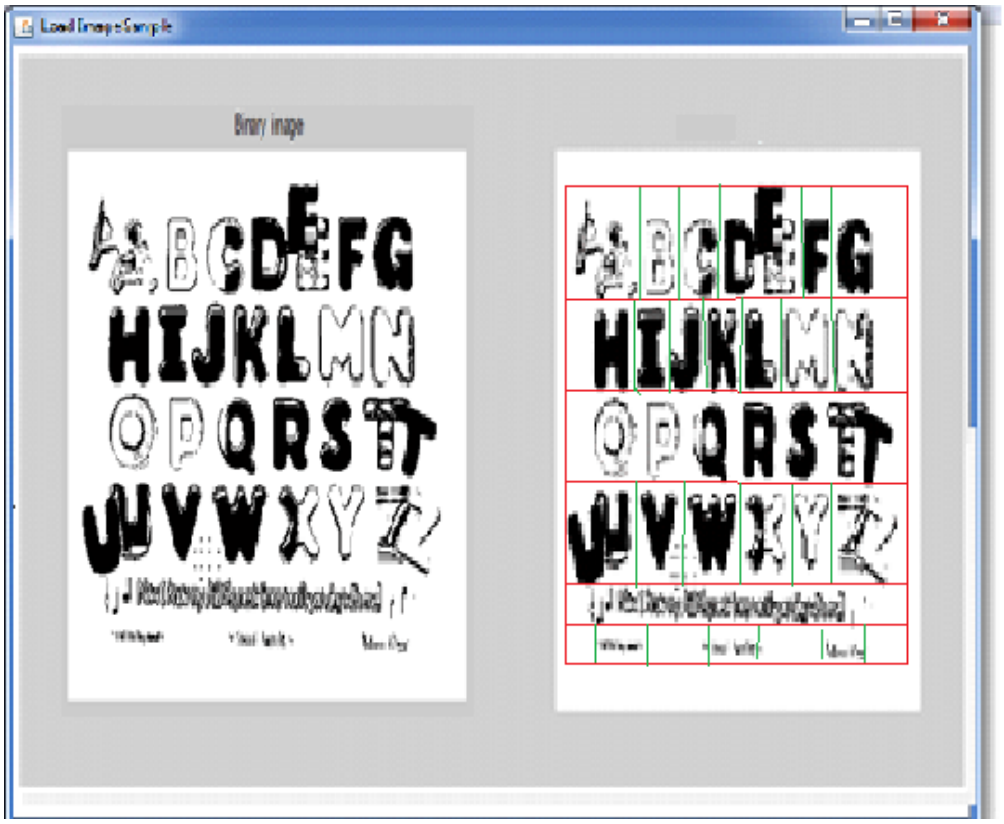


Figure 5.11: Segmentation of Figure 5.4

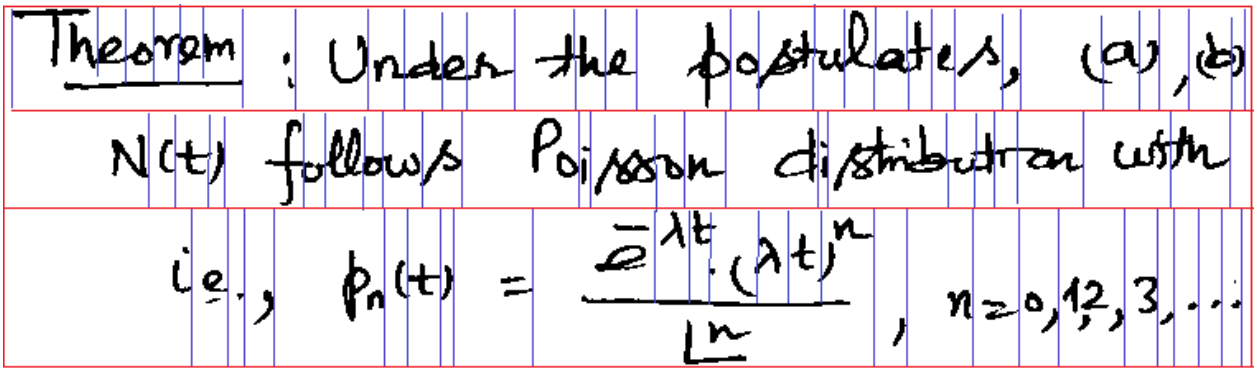


Figure 5.12: Segmentation of Figure 5.7

Here in Figure 5.11 the Left size image is the binarized image which is the output of binarization which is mentioned in Figure 5.4 and right side image is the image after segmentation. Here line segmentation, word segmentation and character segmentation has been done which has been used in the next Recognition Phase. In the same way figure 5.11 is the segmentation phase output of figure 5.7.

5.5 Size Normalization Result

As discussed in section 4.3 Normalization has three types which are as mentioned hereafter; Horizontal Normalization, Vertical Normalization and Generic Normalization.

5.5.1 Horizontal Normalization

In the selection Form as shown in Figure 5.10, on selecting the radio button size Normalization and then clicked on the NEXT link. Then a new Window has been opened which has also three choices. Horizontal Normalization has been selected to normalize the character horizontally. Then it has clicked on the upload button to upload the image.

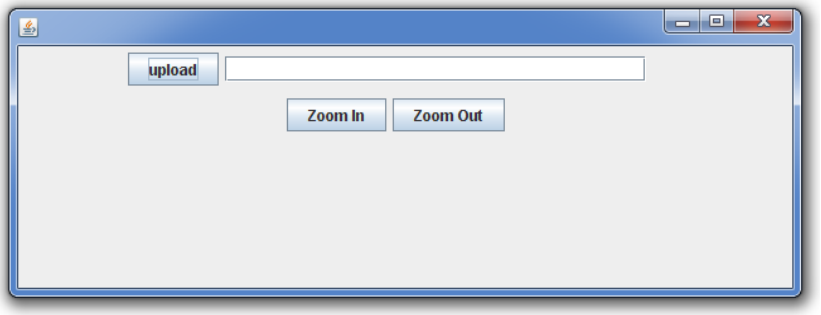


Figure 5.13: Upload Image

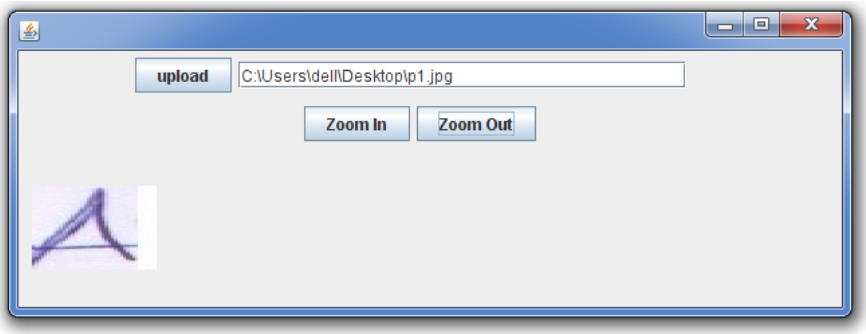


Figure 5.14: Uploaded Original Image for Horizontal Normalization

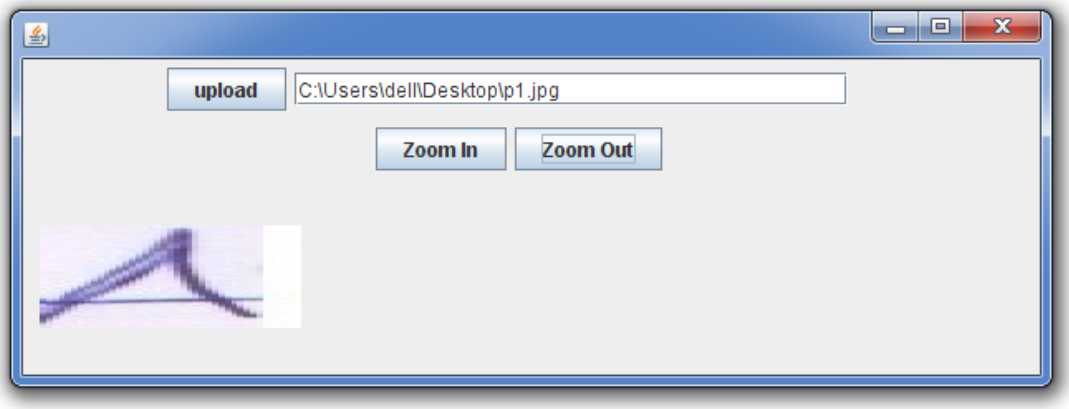


Figure 5.15: Horizontal Normalization

In Figure 5.14 Image has been uploaded to normalize the segmented character horizontally. In Figure 5.15 after clicking on the Zoom in button Normalized image has been obtained according to the requirement.

5.5.2 Vertical Normalization

In the selection Form as shown in Figure 5.10, on selecting the radio button size Normalization and then clicking on the NEXT link, a new Window has been opened which is also having three choices. On Selecting Vertical Normalization here and then clicked on the Next. Then clicked on the upload button to upload the new image as shown in the above Figure 5.13 and 5.14.

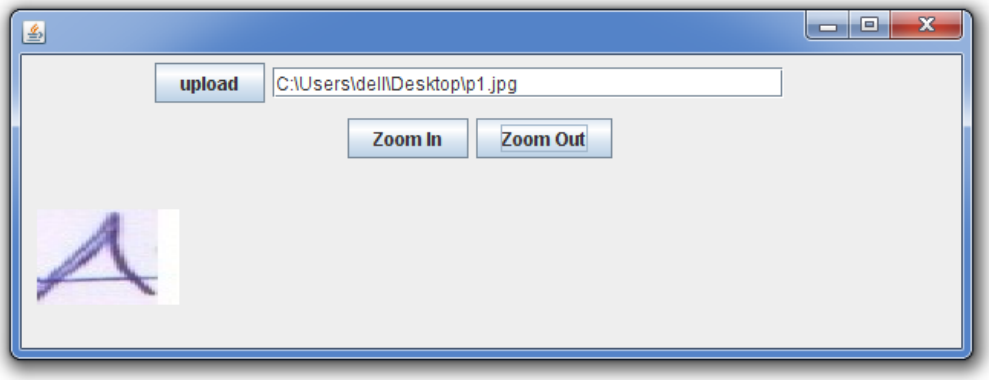


Figure 5.16: Original Image for Vertical Normalization

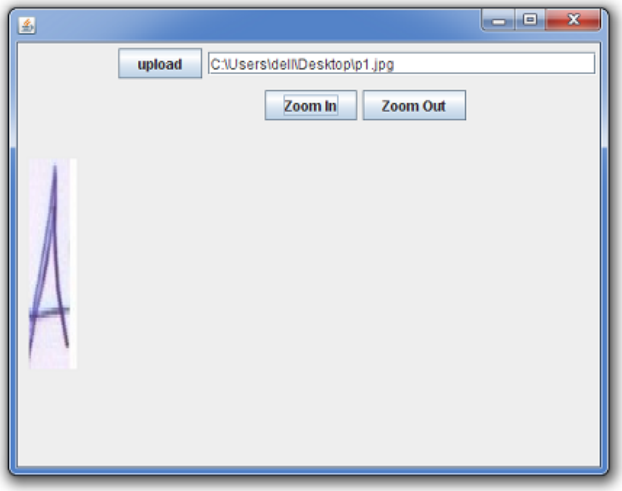


Figure 5.17: Vertical Normalization Zoom In

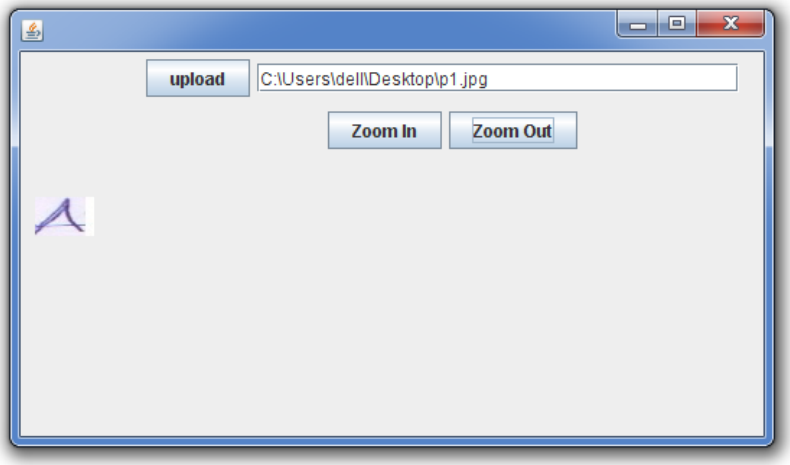


Figure 5.18: Vertical Normalization Zoom Out

In Figure 5.16 Image has been uploaded to normalize the segmented character vertically. In Figure 5.17 after clicking on the Zoom in button Normalized image has been obtained according to the requirement and In Figure 5.18 after clicking on the Zoom out button Normalized image has been obtained according to the requirement. Figure 5.17 and Figure 5.18 is the result of zoom in and zoom out of the individual segmented character of vertical normalization.

5.5.3 Generic Normalization

In the selection Form as shown in Figure 5.10, on selecting the radio button size Normalization and then clicking on the NEXT link, A new Window has been opened which has also three choices. Selected Generic Normalization here and then clicked on the Next.

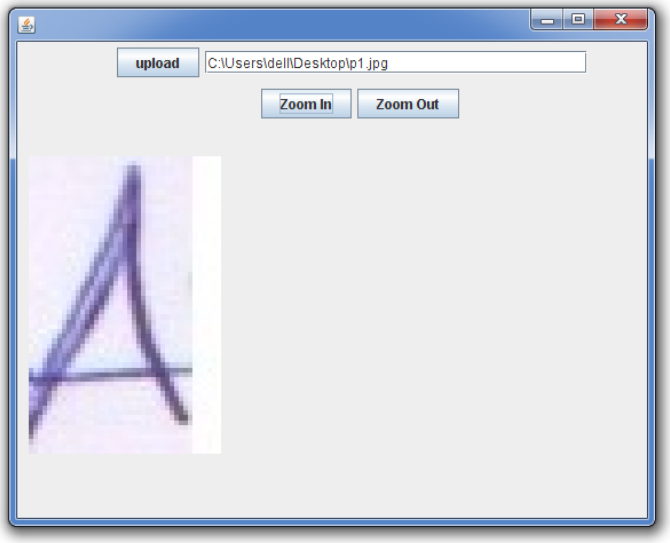


Figure 5.19: Generic Normalization Zoom In

On clicking on the upload button and image has been uploaded as shown in the above Figure 5.13 and Figure 5.14. As shown the original image in Figure 5.16, has been uploaded to normalize the segmented character. In Figure 5.19 after clicking on the Zoom in button Normalized image has been obtained according to the requirement. figure 5.19 is the result of zoom in of the individual segmented character of Generic Normalization.

5.6 Recognition Phase Result

The output of segmented Phase is the input of Numeral Recognition. Segmented characters act as input for the recognition stage. In this phase, firstly the bounding box is created for each character and saved as a separate bitmap file. The bounding box is created around the image of the extracted character using the 4.4 algorithm and the output images are as shown in figure 4.10.

Figure 5.20: Bounding Box of Numerals

The next step is to check each image for the presence of the features, namely, sidebar, loop and water reservoirs. The extracted features are now used by the classification structure to recognize the numerals as discussed in chapter 4. The recognized numerals are stored in an editable form in a document. The explained procedure has been applied to several documents and satisfactory results are obtained by the authors. The results obtained after implementation of all the algorithms are discussed in the following section.

Document	Numerals Present	Numerals Recognized	Accuracy (%)
Doc 1	26	25	96.1
Doc 2	30	28	93.33
Doc 3	12	12	100
Doc 4	17	16	94.11
Doc 5	25	24	96

Table 5.1: Recognition Results for Handwritten English Numerals

The algorithm for the recognition of the handwritten English numerals and algorithm of Segmentation has been applied on a number of documents. The results showing the number of numerals presents in the document and the number of numerals recognized accurately has been tabulated in table number 5.1 and The results showing the number of characters present in the document and the number of segmented character accurately has been tabulated in table number 5.2

Document	Characters Present	Character Segmented	Accuracy (%)
Doc 1	100	95	95
Doc 2	31	28	90.33
Doc 3	12	12	100
Doc 4	10	9	90
Doc 5	25	22	88

Table 5.2: Segmentation Results for Handwritten Characters

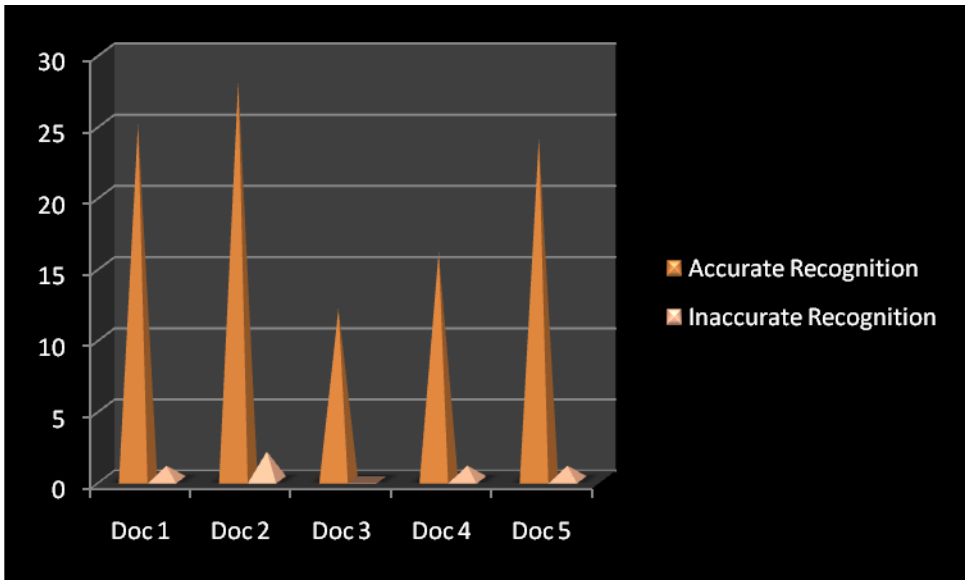


Figure 5.21: Comparison of Accurately Recognized and Inaccurate Recognized Numerals

The comparison between the accurately recognized and inaccurately recognized numerals has been made through graph shown in figure 5.21. The data of table 5.1 has been used here to make this graph, to show the result graphically.

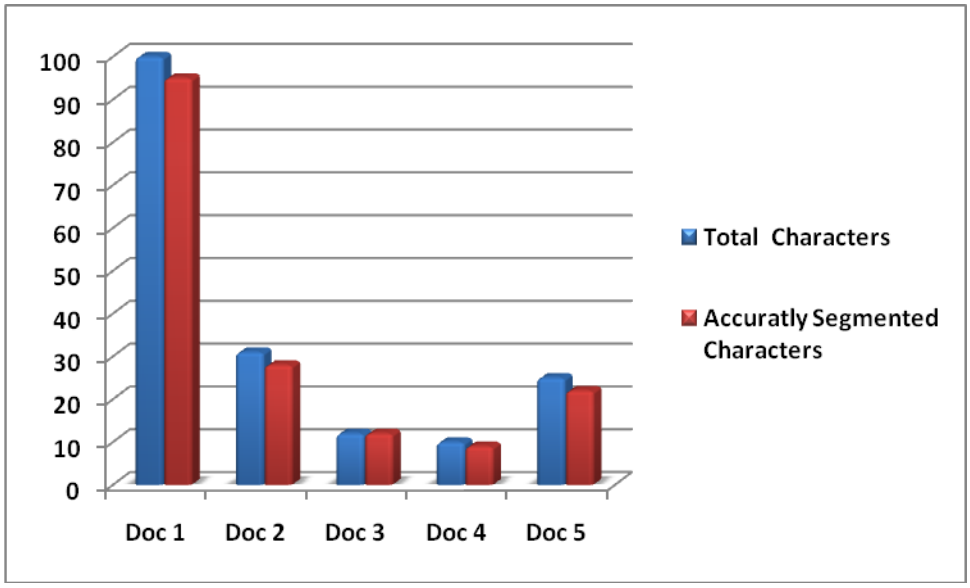


Figure 5.22: Comparison of Accurately Segmented and Total Characters

The comparison between the accurately segmented characters and the total number of characters has been made through graph shown in figure 5.22. The data of table 5.1 has been used here to make the above graph, to show the result graphically.

Thus, the proposed solution implementation results are illustrated through various measures to get a clear picture of the performance of the proposed work.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Character recognition is a process, which associates a symbolic meaning with objects such as letters, symbols and numbers drawn on an image, according to the script used. The practical importance of the OCR applications and the challenging nature of the OCR process have led to huge amount of research in this field. As discussed in Chapter 1, in OCR pre processing phase, first step is scanning that is used to convert a paper document into an image document. The image resulting from the scanning process may contain a certain amount of noise. It may be that the Scanned Document may have varying intensity or skewed. That may affect the OCR system and may cause not accurate segmentation and recognition. Modified average binarization algorithm is more effective and can remove noise and can convert a gray level image into a binary image in those type of image also in which intensity value having a lot of variation. This algorithm results a better binarized image for the OCR system. This acquired image is initially cleaned up using binarization and skew correction. Modified average binarization algorithm overcomes the problem of varying intensity pixel value and skew correction .It will give a better result for the next segmentation phase.

The preprocessed image is segmented using the procedures as discussed in section 4.2. The image is decomposed into lines and lines are further decomposed into word and characters. The results for the segmentation are quite good and enhance the recognition rate.

The segmented numerals are recognized using the technique based on feature extraction and classification. It can be observed from the literature survey that many approaches, which have been developed for the character recognition process, are based on the neural networks, which need high computational power and are testing dependent. Thus, an effort has been made to develop a simple approach for the offline handwritten English numeral recognition. The reorganization technique derives its basis the features extracted for each image is closed loop, sidebars and water reservoirs. Based on these features, a

classification structure is created, which assigns each segmented character to a predefined class. This approach uses features, which are simple to compute. This increases the efficiency of the approach. It does not require much computational power and does not require any kind of training.

The various applications where numeral recognition can be used are reading details like postal zip code, employee code, passport number and processing of forms and bank cheques. It can also be utilized in schools and colleges to process the list of marks. This work can be enhanced to other extents as discussed in the following section.

6.2 Future Scope

This approach can be extended in several ways which are as listed here:-

- The proposed algorithm can be worked upon for the recognition of the English alphabets.
- It is observed that the current approach works for the well-formed characters, so, it can be improvised to work with the degraded documents.
- It can also be improvised to accommodate the variation in the styles of handwriting.
- The classification structure can be used as the basis for the development of similar structures for other Indian languages.
- It can also be improvised to clean up the document which are blurred.
- It can also be used for the documents having slanted words, where slant removal technique can be used in conjunction with the current work.
- It can also need to enhance the performance of existing OCR to get the better result.

REFERENCES

[1] Alginahi, Y., Preprocessing Techniques in Character Recognition. Retrieved 2009**:** http://cdn.intechopen.com/pdfs/11405/InTech/Preprocessing_techniques_in_charactch_recognition.pdf.

[2] Arica, N. and Vural, F. Y. 2001. An Overview of Character Recognition focused on Off-line Handwriting. IEEE Transactions on Systems, *Man, and Cybernetics*, 31 (2), 216-233.

[3] Arora, S., Bhatcharjee, D., Nasipuri, M. and Malik, L., A Two Stage Classification Approach for Handwritten Devanagari Characters in Proceedings of the International Conference on Computational Intelligence and Multimedia Applications, (2007), 377-403.

[4] Banashree, N. P. and Vasantha, R., OCR for Script Identification of Hindi Numerals Using Feature Sub Selection by Means of End Point With Neuromememtic Model in *Proceedings of the World Academy of Science, Engineering and Technology*, (2007), 78-82.

[5] Blumenstein, M., Liu, X. Y. and Verma, N., A Modified Direction Feature for Cursive Character Recognition in *Proceedings of International Joint Conference on Neural Networks*, (2004), 2983-2987.

[6] Blumenstein, M. and Verma, N., A New Segmentation Algorithm for Handwritten Word Recognition in *Proceedings of International Joint Conference on Neural Networks*, (1999), 2893-2898.

[7] Brodic, D., Milivojevic, D. and Tasic, V., Preprocessing of Binary Document Images by Morphological Operator in *Proc. MIPRO, pp*. 883-887,(2011).

[8] Cai, J. and Liu, Z. 1999. Integration of Structural and Statistical Information for Unconstrained Handwritten Numeral Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21 (3), 263-270.

[9] Casey, R. and Lecolinet, A. July 1996. Survey of Methods and Strategies in Character Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, no. 7, pp*. 690-706.

[10] Chang, Z. and Steven, Z. Z .2009. Robust Pre-processing Techniques for OCR Applications on Mobile Devices in *Proc. 6th International Conference on Mobile Technology, Applications, and Systems, Mobility Conference*.

[11] Chou, C. and Wen, H. L. 2010. A binarization method with learning-built rules for document images produced by cameras. *Elsevier Science Inc., Vol. 43, Issue 4, pp.* 1518-1530.

[12] Cuc, D. T. K., Huy, D. Q., An, T. H. and Trong, N. V., Handwritten Number recognition and its Application at Danang University of Technology in *Proceedings of the Conference Report for Scientific Research Students Eighth UD,* (2012).

[13] Ha, T. M. and Bunke, H. 1997. Off-Line, Handwritten Numeral Recognition by Perturbation Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (5), 535-539.

[14] Jayarathna, U. K. S. and Bandara, G. E. M. D. C., New Segmentation Algorithm for Offline Handwritten Connected Character Segmentation in *Proceedings of the First International Conference on Industrial and Information Systems*, (2006), 540-546.

[15] Kasturi, R., Gorman, L. and Govindaraju, V. 2002. Document Image Analysis: A Primer. *Sadhana*, 27 (1), 3–22.

[16] Kaur, A., Kumar, R., and Singh, A. 2010. A Hybrid Approach to Classify Gurmukhi Script Characters. *International Journal of Recent Trends in Engineering and Technology*, 3 (2), 103-105.

[17] Kavallieratou, E. 2000. Slant estimation algorithm for OCR systems. *Wire Communications Laboratory, Electrical Computer Engineering, University of Patras.*

[18] Lehal, G. S. and Singh, C., A Gurmukhi Script Recognition System. in *Proceedings of the Fifteenth International Conference on Pattern Recognition, IEEE Computer Society Press*, (2000), 557-560.

[19] Lu, Y. 1995. Machine Printed Character Segmentation-An Overview. *Pattern Recognition*, 28 (1), 67-80.

[20] Mori, S., Suen, C. Y. and Yamamoto, K., Historical Overview of OCR Research and Development in *Proceedings of the IEEE*, (1992), 1029-1058.

[21] Morita, M. and Bortolozzi, F. 1998. Morphological Approach of Handwritten Word Skew Correction. *IEEE Computer Society Washington*.

[22] O'Gorman, L. and Kasturi, R .1997. Document Image Analysis. *IEEE Computer Society Press.*

[23] Pal, U., Sinha, S. and Chaudhuri, B. B., Multi-Script Line identification from Indian Documents in *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, (2003), 880-884.

[24] Pal, U., Sharma, N., Wakabayashi, T. and Kimura, F. 2008. Handwritten Character Recognition of Popular South Indian Scripts. *Springer Verlag book on Arabic and Chinese Handwriting Recognition*, 251-264.

[25] Patnaik, T. and Gupta, S. 2011. Comparison of Binarization Algorithm in Indian Language OCR. *Proc. IEEE Conference on Language Engineering (LEC'02).*

[26] Ramamurthy, O. V., Roy, S., Narang, V. and Hanmandlu, M. 2012. Devanagari Character Recognition in the Wild. *International Journal of Computer Science*, 38 (4), 38-45.

[27] Ramteke, R. J. and Mehrotra, S. C. 2008. Recognition of Handwritten Devnagari Numerals. *International Journal of Computer Processing of Oriental Languages, Chinese Language Computer Society & World Scientific Publishing Company*.

[28] Singh, R. and Yadav, C. S. 2010. Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network. *International Journal of Computer Science & Communication*, Vol. 1, pp. 91-95.

[29] Stevens, M. E. 1961. Automatic Character Recognition - A State of the Art Report. *National Bureau of Standards Tech Note-112*.

[30] Srihari, S. N. and Lam, S. W. 1995. Character Recognition. Technical Report, CEDAR-TR-95-1.

[31] Syed, B. B., Faisal, S., Thomas, M. B. 2009. Adaptive Binarization of Unconstrained Hand-Held Camera-Captured Document Images. *Journal of Universal Computer Science*, Vol.15, pp. 3343-3363.

[32] Trier, O. D., Jain, A. K. and Taxt, T. 1996. Feature Extraction Methods for Character Recognition - A Survey. *Pattern Recognition*, 29 (4), 641-662.

[33] Zhang, T. Y. and Suen, C. Y. 1984. A Fast Parallel Algorithm for Thinning Digital Patterns. *Communications of the ACM*, 27 (3), 236-239.