Chapter 1
Introduction

Truth is what stands the test of experience.

Albert Einstein
1.1 Introduction to Optical Character Recognition (OCR)

The OCR based technology allows a machine to automatically recognize characters through an optical mechanism. Human beings recognize many objects in this manner as eyes are the biological optical device. The brain interprets the input received from the eyes. The ability to comprehend these signals varies in each person according to various factors [1]. By reviewing these variables, it is possible to understand the challenges associated to this technology. Some examples illustrating the complexity of technologies are mentioned below.

First complexity may arise due to mixing of two different types of characters e.g. mixing English numerals with Russian characters. If we read such kind of documents not knowing one language but familiar with numerals, we may recognize the various characters, but be unable to recognize words. However, on the same page, we are usually able to interpret numerical statements - the symbols for numbers are universally used. This explains why many OCR systems recognize numbers only, [2] while relatively few understand the full alphanumeric character range.

Second, there is similarity between many numerical and alphabetical symbol shapes [3]. For example, while examining a string of characters combining letters and numbers, there is very little visible difference between a capital letter "O" and the numeral "0." As humans, we can re-read the sentence or entire paragraph to help us determine the accurate meaning. This procedure, however, is much more difficult for a machine.

Third, we rely on contrast to help us recognize characters. We may find it very difficult to read text which appears against a very dark background, or is printed over other words or graphics. Again, programming a system to interpret only the relevant data and disregard the rest is a difficult task for OCR engineers.

There are many other problems which challenge the developers of OCR systems. In this thesis, we will review the history, advancements, abilities and limitations of existing systems.
The Optical recognition system will be either off line or on-line. The off-line system is done after the writing or printing has been completed, whereas on-line recognition the computer recognizes the characters as they are drawn [4]. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents. The figure 1 illustrates different areas of the OCR.

In general the more constrained the input is, the better performance of the OCR system is expected using knowledge of constrains. However, when it comes to totally unconstrained handwriting, OCR machines are still a long way from reading as well as humans [5]. However, the potential for technologies is tremendous. Attempts are in progress to bring unconstrained OCR close to ideal, or better than ideal.

Figure 1.1: The different areas of character recognition.
1.2 History of OCR

The engineering attempts at automated recognition of printed characters started prior to World War II. But it was not until the early 1950s that a commercial venture was identified that justified necessary funding for research and development of the technology. This impetus was provided by the American Bankers Association and the Financial Services Industry [6]. They challenged all the major equipment manufacturers to come up with a "Common Language" to automatically process checks. After the war, check processing had become the single largest paper processing application in the world. Although the banking industry eventually chose Magnetic Ink Recognition (MICR), some vendors had proposed the use of an optical recognition technology. However, OCR was still in its infancy at the time and did not perform as acceptably as MICR. The advantage of MICR was that it is relatively impervious to change, fraudulent alteration and interference from non-MICR inks.

The "eye" of early OCR equipment utilized lights, mirrors, fixed slits for the reflected light to pass through, and a moving disk with additional slits. The reflected image was broken into discrete bits of black and white data, presented to a photo-multiplier tube, and converted to electronic bits.

The "brain's" logic required the presence or absence of "black" or "white" data bits at prescribed intervals. This allowed it to recognize a very limited, specially designed character set.

To accomplish this, the units required sophisticated transports for documents to be processed. The documents were required to run at a consistent speed and the printed data had to occur in a fixed location on each and every form.

The next generation of equipment, introduced in the mid to late 1960's, used a cathode ray tube, a pencil of light, and photo-multipliers in a technique called "curve following". These systems offered more flexibility in both the location of the data and the font or design of the characters that could be read. It was this technique that introduced the
concept that handwritten characters could be automatically read, particularly if certain constraints were utilized [7]. This technology also introduced the concept of blue, on-reading inks as the system was sensitive to the ultraviolet spectrum.

The third generation of recognition devices, introduced in the early 1970's, consisted of photo-diode arrays. These tiny little sensors were aligned in an array so the reflected image of a document would pass by at a prescribed speed [8]. These devices were most sensitive in the infra-red portion of the visual spectrum so "red" inks were used as non-reading inks. That brings us to this generation of hardware.

**A Brief History of Printed Script OCR**

- **1870** Carey invented Retina Scanner for image transmission using Mosaic of Photo cells.
- **1900** Tyurin attempted OCR as reading aid for the blind.
- **1929** Tausheck obtained first OCR patent in Germany.
- **1950** Dr. Shepard from Intelligent Machine Research Co developed a workable OCR.
- **1957** Electronic Reading Automation made by Solartron Electronics Ltd. was announced.
- **1965** IBM OCR system IBM 1287 was exhibited in New-York fair.
- **2006** Many excellent systems for Roman based scripts (Scansoft Omnipage, Abby Fine-reader, Charactell Softwriting, Dan Ching Penpower) are available in the market.
The history of handwriting recognition systems is not complete without mentioning the Optical Character Recognition (OCR) systems which preceded them. Optical Character Recognition (OCR) is a problem recognized as being as old as the computer itself [9], [10]. There have been many papers and technical reports published reviewing the history of OCR technologies [11], [12]. Modern OCR was said to have begun in 1951 due to an invention by M. Sheppard called GISMO, a robot reader-writer [13]. In 1954, a prototype machine developed by J. Rainbow was used to read uppercase typewritten letters at very slow speeds. By 1967, companies such as IBM finally marketed OCR systems. However in the late 60’s, these systems were still very expensive, and therefore could only be used by large companies and government agencies. Today, OCR systems are less expensive and can recognize more fonts than ever before. Even so it is important to note that in some situations these commercial packages are not always satisfactory. Senior [14] mentions that problems still exist with unusual character sets, fonts and with documents of poor quality. Research now focuses more on hand-printed numeral, character and joined /cursive handwriting recognition.

Unfortunately the success of OCR could not carry on to handwriting recognition, due to the variability in people’s handwriting. As for the recognition of isolated handwritten numerals, Suen [15], details many researchers which have already obtained very promising results using various classification methods. Suen mentions that the key to high recognition rates is feature extraction [16]. However, this in itself is a very difficult problem which has led researchers to use more complex methods for preprocessing, feature extraction and classification. Such methods include the use of Neural Networks and Mathematical Morphology.

**Recent Achievements**

Researchers all over the world have achieved successful results in handwriting recognition [17], [18], [19], [20]. We present some of these results below in Table 1. As can be seen the table is divided into 3 main categories: handwritten numeral recognition, character and cursive word recognition.
<table>
<thead>
<tr>
<th>Authors</th>
<th>Recognition Rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeral</td>
<td></td>
</tr>
<tr>
<td>Denker et al. [21]</td>
<td>86</td>
</tr>
<tr>
<td>Bottou et al.[22]</td>
<td>91.9</td>
</tr>
<tr>
<td>Srihari [23]</td>
<td>89-93</td>
</tr>
<tr>
<td>Lee[24]</td>
<td>99.5</td>
</tr>
<tr>
<td>Character</td>
<td></td>
</tr>
<tr>
<td>Liou and yang[26]</td>
<td>88-95</td>
</tr>
<tr>
<td>Srihari [27]</td>
<td>85-93</td>
</tr>
<tr>
<td>Shustorovich[28]</td>
<td>89.40-96.44</td>
</tr>
<tr>
<td>Cursive word</td>
<td></td>
</tr>
<tr>
<td>Edelman et al.[29]</td>
<td>50</td>
</tr>
<tr>
<td>Lecolinet et al.[30]</td>
<td>53</td>
</tr>
<tr>
<td>Leroux et al.[31]</td>
<td>62</td>
</tr>
<tr>
<td>Chen et al.[32]</td>
<td>64.9-72.3</td>
</tr>
<tr>
<td>Bozinovic et al.[33]</td>
<td>54-72</td>
</tr>
<tr>
<td>Senior and Fallside[34]</td>
<td>78.6</td>
</tr>
<tr>
<td>Simon[35]</td>
<td>86</td>
</tr>
<tr>
<td>Guillevic and Suen[36]</td>
<td>76-98.5</td>
</tr>
<tr>
<td>Bunke et al.[37]</td>
<td>98.27</td>
</tr>
</tbody>
</table>

In Numeral recognition there are many excellent results, one of the best obtained by Lee [38]. Also in Cursive word recognition the same high standard of results may be found, Bunke et al. [39] obtaining the highest recognition accuracy. Unfortunately, as previous reviewers have mentioned, there are too many factors which do not allow a suitable
comparison to be performed. The influencing factors relate directly to the difference in conditions for experimentation. One of the main differences is the type of handwriting database used for experimentation. In some cases, researchers have constrained their experiments heavily, only using one person’s handwriting, while other researchers’ experiments were not performed on benchmark databases. Some of the problems and challenges which are faced by researchers today include: developing accurate segmentation, preprocessing, feature extraction and classification techniques.

For the first problem, segmentation, the diverse styles and sizes of handwriting both play a large factor in the failure of current techniques. In some cases even a human being would not be able to segment handwriting containing characters which are tightly packed together and illegible [40]. These segmentation systems also have to deal with the variability of handwriting from one person to another, not to mention problems when one writer’s handwriting is cursive, while another person’s is simply overlapping. Challenges faced for preprocessing deal with the choice of whether to convert raw handwriting into a more efficient form i.e. whether to binaries the handwriting or keep it in grey-scale form. Other researchers are disputing whether the handwriting should be thinned or should remain the way it is to preserve features.

Feature extraction poses the problem of choosing the right features to extract and the right technique to perform the task [41]. For example researchers may choose between extracting features such as the entire contours of characters or by extracting many features such as end-points, loops, holes and so on. Finally, the task of finding a suitable classification technique (for individual characters and whole words) has been exhaustively pursued. However, again the variability of handwriting and the lack of reliable feature extraction and preprocessing techniques have impaired many unconstrained approaches. For most of the aforementioned problems, including feature extraction and classification, researchers have turned to complex and intelligent methods. The use of Neural Networks has become extremely popular, and may hopefully enable researchers to move closer to solving the handwriting recognition problem.
1.3 Challenges and applications

There are various types of documents that can be digitized. For example, books, letters, bank checks, handwritten texts, shorthand, musical score papers, maps, drawings, diagrams, mathematical expressions, web documents, text in photographs.

There is a widespread appearance of commercial optical character recognition products meeting the requirements of different users. In this chapter we treat some of the different areas of application for OCR. Three main application areas are commonly distinguished; data entry, text entry and process automation. Over time, other applications evolved.

Data entry

This area covers technologies for entering large amounts of restricted data. Initially such document reading machines were used for banking applications. The systems are characterized by reading only an extremely limited set of printed characters, usually numerals and a few special symbols. They are designed to read data like account numbers, customer’s identification, article numbers, amounts of money etc. [42] The paper formats are constrained with a limited number of fixed lines to read per document. Because of these restrictions, readers of this kind may have a very high throughput of up to 150,000 documents per hour. Single character error and reject rates are 0.0001% and 0.01% respectively. Also, due to the limited character set, these readers are usually remarkably tolerant to bad printing quality. These systems are specially designed for their applications and prices are therefore high.

Text entry

The second branch of reading machines is that of page readers for text entry, mainly used in office automation. Here the restrictions on paper format and character set are exchanged for constraints concerning font and printing quality. The reading machines are used to enter large amounts of text, often in a word processing environment. These page readers are in strong competition with direct key-input and electronic exchange of data.
This area of application is therefore of diminishing importance [43]. As the character set read by these machines is rather large, the performance is extremely dependent on the quality of the printing. However, under controlled conditions the single character error and reject rates are about 0.01% and 0.1% respectively. The reading speed is typically in the order of a few hundred characters per second the word, by saying for instance that after a period there should usually be a capital letter. Also, for different languages the probabilities of two or more characters appearing together in a sequence can be computed and may be utilized to detect errors. For instance, in the English language the probability of a “k” appearing after an “h” in a word is zero and if such a combination is detected an error is assumed.

Another approach is the use of dictionaries, which has proven to be the most efficient method for error detection and correction. Given a word, in which an error may be present, the word is looked up in the dictionary. If the word is not in the dictionary, an error has been detected, and may be corrected by changing the word into the most similar word. Probabilities obtained from the classification, may help to identify the character which has been erroneously classified [44]. If the word is present in the dictionary, this does unfortunately not prove that no error occurred. An error may have transformed the word from one legal word to another, and such errors are undetectable by this procedure. The disadvantage of the dictionary methods is that the searches and comparisons implied are time-consuming.

**Automatic cartography**

Character recognition from maps presents special problems within character recognition. The symbols are intermixed with graphics, the text may be printed at different angles and the characters may be of several fonts or even handwritten.

**Form readers**

Such systems are able to read specially designed forms. In such forms all the information irrelevant to the reading machine is printed in a color “invisible” to the scanning device. Fields and boxes indicating where to enter the text is printed in this invisible color.
Characters should be entered in printed or handwritten upper case letters or numerals in the specified boxes [45]. Instructions are often printed on the form as how to write each character or numeral. The processing speed is dependent on the amount of data on each form, but may be about a few hundred forms per minute. Recognition rates are seldom given for such systems.

**Process automation**

Within this area of application the main concern is not to read what is printed, but rather to control some particular process. This is actually the technology of automatic address reading for mail sorting. Hence, the goal is to direct each letter into the appropriate bin regardless of whether each character was correctly recognized or not. The general approach is to read all the information available and use the postcode as a redundancy check. The acceptance rate of these systems is obviously very dependent on the properties of the mail. This rate therefore varies with the percentage of handwritten mail. Although, the reject rate for mail sorting may be large, the miser rate is usually close to zero. The sorting speed is typically about 30,000 letters per hour.

**Other applications**

The above areas are the ones in which OCR has been most successful and most widely used. However, many other areas of applications exist, and some of these are mentioned below.

**Aid for blind**

In the early days, before the digital computers and the need for input of large amounts of data emerged, this was the imagined area of application for reading machines. Combined with a speech synthesis system such a reader would enable the blind to understand printed documents [46]. However, a problem has been the high costs of reading machines, but this may be an increasing area as the costs of microelectronics fall.
Automatic number-plate readers.

A few systems for automatic reading of number plates of cars exist. As opposed to other applications of OCR, the input image is not a natural bi-level image, and must be captured by a very fast camera. This creates special problems and difficulties although the character set is limited and the syntax restricted [47]. They included cash register tape readers, page scanners, etc. Any standard form or document with repetitive variable data would be a candidate application for OCR. Some very imaginative applications have evolved. Perhaps the most innovative are the Kurzwell scanners which read for the blind. With these devices, the optically scanned pages are converted to spoken words.

1.4 Limitations

OCR has never achieved a read rate that is 100% perfect. Because of this, a system which permits rapid and accurate correction of rejects is a major requirement. Exception item processing is always a problem because it delays the completion of the job entry, particularly the balancing function [48]. Of even greater concern is the problem of misreading a character (substitutions). In particular, if the system does not accurately balance dollar data, customer dissatisfaction will occur. The success of any OCR device to read accurately without substitutions is not the sole responsibility of the hardware manufacturer. Much depends on the quality of the items to be processed. Through the years, the desire has been [49]:

- to increase the accuracy of reading, that is, to reduce rejects and substitutions
- to reduce the sensitivity of scanning to read less-controlled input
- to eliminate the need for specially designed fonts (characters), and
- to read handwritten characters.

However, today's systems, while much more forgiving of printing quality and more accurate than earlier equipment, still work best when specially designed characters are used and attention to printing quality is maintained. However, these limits are not objectionable to most applications, and dedicated users of OCR systems are growing each year. But the ability to read a special character is not, by itself, sufficient to create a successful system.
1.5. Successful OCR System.

1. It takes a complimentary merging of the input document stream with the processing requirements of the particular application with a total system concept that provides for convenient entry of exception type items with an output that provides cost effective entry to complete the system. To show a successful example, let's review the early credit card OCR applications. Input was a carbon imprinted document. However, if the carbon was wrinkled, the imprinter was misaligned, or any one of a variety of reasons existed, the imprinted characters were impossible to read accurately.

2. To compensate for this problem, the processing system permitted direct key entry of the fail to read items at a fairly high speed. Directly keyed items from the misread document were under intelligent computer control which placed the proper data in the right location for the data record. Important considerations in designing the system encouraged the use of modulus controlled check digits for the embossed credit card account number. This, coupled with tight monetary controls by batch totals, reduced the chance of read substitutions.

3. The output of these early systems provided a "country club" type of billing. That is, each of the credit card sales slips was returned to the original purchaser. This provided the credit card customer with the opportunity to review his own purchases to insure the final accuracy of billing. This has been a very successful operation through the years. Today's systems improve the process by increasing the amount of data to be read, either directly or through reproduction of details on the sales draft. This provides customers with a "descriptive" billing statement which itemizes each transaction.
OCR SYSTEMS

OCR systems may be subdivided into two classes. The first class includes the special purpose machines dedicated to specific recognition problems. The second class covers the systems that are based on a PC and a low-cost scanner.

1.5.1 Dedicated hardware systems

The first recognition machines were all hardwired devices. Because this hardware was expensive, throughput rates had to be high to justify the cost, and parallelism was exploited. Today such systems are used in specific applications were speed is of high importance, for instance within the areas of mail-sorting and check-reading. The costs of these machines are still high, up to a million dollars, and they may recognize a wide range of fonts.

1.5.2. Software based PC versions

Advancements in the computer technology have made it possible to fully implement the recognition part of OCR in software packages which work on personal computers. Present PC systems are comparable to the large scaled computers of the early days, and as little additional equipment is required, the costs of such systems are low. However, there are some limitations in such OCR software, especially when it comes to speed and the kinds of character sets read. Hand held scanners for reading do also exist. These are usually limited to the reading of numbers and just a few additional letters or symbols of fixed fonts [50]. They often read a line at a time and transmit it to application programs. Three commercial software products are dominant within the area of recognition of European languages. These are systems produced by Caera Corporation, Kurzweil and Calera Corporation, with prices in the range of $500 - $1000. The speed of these systems is about 40 characters per second.
Each step requires special attention to make OCR successful:

Input

When installing an OCR system, the most important consideration is the manner of creating input.

1. How do you intend to create the input? If the input is typewritten data, how many different typewriters will create the input? Will they be electronic, electric, or manual? What type styles or fonts do they have? Will the typewritten material be from a fabric or carbon ribbon? This gives you an idea of the information you need to obtain.

2. What kind of a document will be used for the application? For most systems, the data to be scanned must occur in the same location from document to document. Guide lines or the location of data identifiers, need to be pre-printed. Do they need to be in a "non-reading" color (drop-out ink)? Where will they be printed? What size will they be? Will the form meet the requirements specified by the scanner manufacturer? Will the right data be in the right location for best digit or balancing routines to facilitate performance? Remember, your attention to detail and reviewing the "what if" possibilities before installation will save a tremendous amount of dissatisfaction later.

3. How will input be handled both prior to preparation, after printing, and after processing? If accurate registration must be maintained, moisture-proof wrapping your pre-printed forms may be necessary. If the item is to be mailed to the processing center individually, you may want to prescribe a heavy duty envelope to prevent damage in transit. If the items are to be picked up in large quantities, a special basket or other carrier may be required to ensure documents are not damaged. Although a rubber band is a fine tool to bind a group of documents together, it is a prime cause of damage to paper documents that are to be processed in an automatic feeding device. If you require subsequent archival of the documents for retrieval purposes, proper storage containers are required. You may also need a pre-printed serial number to help research archived material.
These are but a few of the questions that need to be answered. Your OCR system manufacturer is the best source of information on individual system input requirements.

**OCR capabilities**

The sophistication of the OCR system depends on the type and number of fonts recognized. Below a classification, by the order of difficulty, based on the OCR systems’ capability to recognize different character sets, is presented.

**Fixed font**

OCR machines of this category deals with the recognition of one specific typewritten font. Such fonts are OCR-A, OCR-B, Pica, Elite, etc. These fonts are characterized by fixed spacing between each character. The OCR-A and OCR-B are the American and European standard fonts specially designed for optical character recognition, where each character has a unique shape to avoid ambiguity with other characters similar in shape.

Using these character sets, it is quite common for commercial OCR machines to achieve a recognition rate as high as 99.99% with a high reading speed. The systems of the first OCR generation were fixed font machines, and the methods applied were usually based on template matching and correlation.

**Multifont**

Multifont OCR machines recognize more than one font, as opposed to a fixed font system, which could only recognize symbols of one specific font. However, the fonts recognized by these machines are usually of the same type as those recognized by a fixed font system. These machines appeared after the fixed-font machines. They were able to read up to about ten fonts. The limit in the number of fonts were due to the pattern recognition algorithm, template matching, which required that a library of bit map images of each character from each font was stored. The accuracy is quite good, even on degraded images, as long as the fonts in the library are selected with care.
Omni font

An omni font OCR machine can recognize most no stylized fonts without having to maintain huge databases of specific font information. Usually omnifont-technology is characterized by the use of feature extraction. The database of an omnifont system will contain a description of each symbol class instead of the symbols themselves. This gives flexibility in automatic recognition of a variety of fonts. Although omnifont is the common term for these OCR systems, this should not be understood literally as the system being able to recognize all existing fonts. No OCR machine performs equally well or even usably well, on all the fonts used by modern typesetters. A lot of current OCR-systems claim to be omnifont.

Constrained handwriting

Recognition of constrained handwriting deals with the problem of unconnected normal handwritten characters. Optical readers with such capabilities are not yet very common, but do exist. However, these systems require well-written characters, and most of them can only recognize digits unless certain standards for the hand-printed characters are followed see figure 2. The characters should be printed as large as possible to retain good resolution, and entered in specified boxes. The writer is also instructed to keep to certain models provided, avoiding gaps and extra loops. Commercially the term ICR (Intelligent Character Recognition) is often used for systems able to recognize hand printed characters. And the IEEE fellows surveyed is given in figure 3
### OCR COMPLETION GUIDANCE

**RULES**

1. Use black pen whenever possible.

2. Form large characters, but within the box edges.

3. Use simple shapes, avoid loops or curls or flourishes.


5. Connect lines.

6. Do not use alternative shape four continental seven continental one.

7. Do not link characters.

8. Do not overlap characters.

**EXAMPLES**

<table>
<thead>
<tr>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>32460</td>
<td>32460</td>
</tr>
<tr>
<td>23705</td>
<td>23705</td>
</tr>
<tr>
<td>0689</td>
<td>0689</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>471</td>
<td>471</td>
</tr>
<tr>
<td>5621</td>
<td>5621</td>
</tr>
<tr>
<td>4762</td>
<td>4762</td>
</tr>
</tbody>
</table>

**Alpha Character Set**

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

**Numeric Character Set**

1 2 3 4 5 6 7 8 9 0

---

Figure 1.2: Instructions for OCR handwriting.
Summary

Character recognition techniques associate a symbolic identity with the image of character. Character recognition is commonly referred to as optical character recognition (OCR), as it deals with the recognition of optically processed characters. The modern version of OCR appeared in the middle of the 1940’s with the development of the digital computers. OCR machines have been commercially available since the middle of the 1950’s. Today OCR-systems are available both as hardware devices and software packages, and a few thousand systems are sold every week.
1.6. Motivation and objective

Character recognition is front line research problem that has been going on for last five decades. Nonetheless, it still is an active area of research because the problem is complex in nature and full of trade-offs. No solution has been offered that solves the problem both efficiently and completely.

The problem consists of recognizing which character is represented by a given set of points. It has been argued that, while this problem is complex it is not fundamental solving character recognition and is primarily a practical problem. This is echoed in the following quote from Jean Ward [51]:

“An open question now, after over 30 years of published handwriting recognition research and development, is exactly what useful purpose our work will serve. There are very few, if any, successful applications using handwriting recognition.”

As has been mentioned previously, the recognition of handwritten words via computers is a problem that may be traced to the first quarter of 20\textsuperscript{th} centuries. The recognition accuracy for separated handwritten numerals and characters has improved significantly in resent years. Many commercial and accurate systems are available now [52]. Unfortunately the success obtained on machine printed OCR has not readily transferred to the handwritten recognition arena. This may be attributed to the immense variation in human handwriting [53]. However the final frontier remains the accurate recognition of handwritten printed and cursive words. Again research in this area has also been phenomenal, many systems currently exist. [54], [55]. However these are mostly worked on English, Latin, Chinese scripts. Even some research is going on Devangari script, Telugu and some regional Indian Languages, but the pursuit of more accurate recognition rates continues to spur in this field. It can also be seen that along with the challenging nature of handwritten word recognition problem, immense potential lies in the commercial sector to apply the system created.

Major research works in Indian languages are happening at CDAC GIST group, ISI Kolkata, IIIT Hyderabad and Department of computer Science and Engineering,
University of Buffalo. Motivation for this research is due to the challenges in Indian languages, tremendous number of application and the development of algorithms for intelligence of the system by mimic the human intelligence.

1.7. Significance of the Thesis

This area of research is an important in recent years or we can say now days this is burning topic of research. Lot of research work has done in this area but still the problem are exists relate to performance evolution and robustness of character recognition system . In this proposed research work we have presents different methods as well as a more through evaluation . The contributions of this thesis are as follows:

This research work is to study principles behind character recognition which includes; the need of the character recognition. Digital document processing is gaining popularity for application to office & library automation, bank & postal services, publishing & communication technology.

Further the focus is given on the two distinct areas of research concerning Handwritten Character Recognition, Off-line Character recognition and On-line Character Recognition. Off-line character description data only provides the list of coordinates. This is due to the reason that these images are usually of characters that were written earlier and later converted to digital format using a digital scanner. In contrast, On-line character description data provides much more information for perception. Features like, the direction of movement, the number of pen-ups and pen downs etc. are extremely useful. Also, the coordinates are presented in a time-ordered fashion, which allow the system to know the order of the coordinates and segments in which the writer wrote the character. The challenge with On-line character recognition is the development of a system that can recognize these Characters in real-time.

Recognition of printed English numerals has been a popular research area because of its various possible applications. Some of its application areas are automatic postal sorting, bank cheque processing, form processing etc.
In our work, A Block Based approach is used for Recognition of printed numerals with different font size. A block based technique is applied on 5 different font sizes viz. 6, 8, 10, 12 and 14 on numerals with 0 to 9. A linear scaling process is used for Normalization. The information has been extracted by sorting regions in order of decreasing density. Out of which, 10 density features were selected for recognition. Minimum distance classifier has been adopted for recognition. Its performance was found to be satisfactory.

The neural network technology is also used to enhance recognition rate from features extracted from block based technique. The training of NN has been started from 700 epochs and reached to 2,28,000 epochs to get 100% Recognition Rate. Time constrain is also discussed in detail. The optimization of the network design is achieved for the problem.

The main thrust of this thesis considers problems related to online recognition. In this method, velocity vectors of pen movement are input features. Online features of characters are extracted and used for recognition purpose.

At last the Analysis, Interpretation conclusion & future work of the research topic is given. This work refers on adequate number of well-reputed reference papers.

1.8. Outline of the Thesis

The other chapters of the thesis are organized as follows. Chapter 2 deals with Automatic Character Recognition System. Systems are classified according to the text type and data acquisition techniques. In chapter 3, the recognition of printed English numerals irrespective of their font size. A block based technique is applied on 5 different font sizes viz. 6, 8, 10, 12 and 14 on numerals with 0 to 9. Different structures of Neural Network & Recognition of digits by Neural Network are discussed in chapter 4. The chapter 5 includes description of work done related on line character recognition. Contributions of the thesis & their salient features and scope of future work are summarized in chapter 6.
References


52. R. J. Ramteke, S. C. Mehrotra, “Feature Extraction Based on Moment Invariants for Handwriting Recognition” at 2006 IEEE International Conference on Cybernetics and Intelligent System (CIS), Bangkok, Thailand on 7th – 9th June 2006. (Finalist for Best Student Paper Award)

