Chapter 1
Introduction

Introduction

1.1 Optical Character Recognition

1.2 Brief History of Character Recognition

1.3 Bangla Scripts for OCR

1.4 Problem Definition

1.5 Background Study

1.6 Objective of the Thesis

1.7 Decomposition of Thesis

Conclusion

Introduction

Human functions like reading are simulated by machine is an ancient dream. However, machine reading has turned into reality over few decades. Optical Character Recognition (OCR) has likewise turned out to be a standout amongst the best uses of innovation in the field of pattern recognition and artificial intelligence. It has begotten an ample field of research in pattern recognition and artificial intelligence [1]. Since the mid-1950s, OCR has been an extremely dynamic field of innovative work. While the OCR technology for some scripts like Latin is fairly mature and commercial OCR systems like Nuance OmniPage Pro or ABBYY FineReader are available which can perform with high accuracy, it is still under development for other scripts like Chinese, Devangari and Bangla.

In spite of the fact that a lot of researches have been found to contemplate on Roman scripts and scripts related to European languages and Asian languages [2]. This issue is more noticeable for different contents for which OCR innovation is relatively newer. Typefaces are very important in determining the performance of the OCR technology. Hence in order to improve the accuracy of the OCR system, typefaces which are specially designed for OCR are required. For Latin content, many typefaces have been composed which are streamlined for OCR. These uncommonly designed typefaces have a novel and very much characterized character set which takes into account more noteworthy exactness in acknowledgment. This in turns helps in building minimal effort frameworks which can perceive characters utilizing basic calculations.

1.1 Optical Character Recognition

Optical Character Recognition, or OCR, is an innovation that empowers to convert to different types of documents, for example, scanned paper documents, PDF files or images captured by a digital camera into editable and accessible information. If one has a paper record - such as, magazine article, handout, or PDF contract sent to him or her by email. Clearly, a scanner is not sufficient to make this data accessible for altering, say in Microsoft Word. All a scanner can do is make a picture or a preview of the report that is nothing more than a collection of black and white or color dots, known as a raster image. In order to separate and repurpose information from scanned documents, camera image or image-only PDFs, an OCR software is required that would single out letters on the image, articulate them and then - words into sentences, subsequently empowering to get access and edit the content of the original document.

1.2 Brief History of Character Recognition

The fantasy of influencing machines to perform humane task like reading is not new. The primary attempt was in 1870 when C. R. Carey imagined an image transmission system. During the first decade of 19th Century many endeavors were made. However, the frontier form of OCR appeared in 1940s.

The first OCR was introduced in Reader's Digest in 1954. It was used to convert typewritten report to punched card so that they can be input in the computer. The first commercial OCR appeared from 1960 to 1965. These OCR had a constrained letter shape read. The characters were specifically designed for machine recognition and were not very natural. With time the OCR was able to recognize up to 10 different fonts. The reading machines of the second generation appeared in the middle of the 1960's and early 1970's. These systems were able to recognize regular machine printed characters and also had hand-printed character recognition capabilities. The first one of this kind was IBM 1287. In this period, characters in Latin script were also standardized. OCR-A and OCR-B were also designed in this period. These fonts were designed so that they can be recognized by a machine but were also still readable by a human. These first appeared in the middle of 1970s. The challenge was to recognize poorly scanned documents ad hand-written character set. Also low cost and high accuracy were main objective which was achieved also because of the advancement in the technology. Today OCRs are available at a very low cost and OCR systems are also available as software package. Omnifont OCRs are available for Latin script. Although systems are available for Latin, Cyrillic, far eastern and many Middle Eastern scripts, such systems for Devanagari are still in the research and development stage. This is mainly due to a lack of a commercial market.

1.3 Bangla Scripts for OCR

Bangla is the national language of Bangladesh and the second most popular language in India. The Bengali (also called Bangla) script is used for writing the Bengali language, spoken by over 180,000,000 people all over the world. It is also used for a number of other Indian languages including Sylheti and, with one or two modifications, Assamese. It is a Brahmic script although its exact derivation is disputed. Bengali writing shares some similarities with the Dravidian-language scripts, particularly in the shapes of some vowel letters, but it is generally more similar to the Aryan-language scripts, in particular Devanagari.

The alphabet of the modern Bangla script consists of 11 vowels and 39 consonants. These characters are called as basic characters. Writing style in Bangla is from left to right and the concept of upper/lower case is absent in this script. It can be seen that most of the characters of Bangla have a horizontal line (Matra) at the upper part. From a statistical analysis we notice that the probability that a Bangla word will have horizontal line is 0.994. In Bangla script a vowel following a consonant takes a modified shape. Depending on the vowel, its modified shape is placed at the left, right, both left and right, or bottom of the consonant. These modified shapes are called modified characters. A consonant or a vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character. Compound characters can be combinations of two consonants as well as a consonant and a vowel. Compounding of three or four characters also exists in Bangla. There are about 280 compound characters in Bangla. To get an idea of Bangla basic characters and their variability in handwriting, a set of handwritten Bangla basic characters are shown in Fig. 1.1.

Fig. 1.1 Basic characters & numerals of Bangla script

Despite the fact that Bangla is an extremely rich and old dialect, the matter of disappointment is its computerization has not yet gone much far. In fact, dedicated thesis and development for Bangla computerization has just been started from the last decade. In computerization of any language, one of the vital tasks is to develop an efficient and effective Optical Character Recognition (OCR)

system for the respected language. In order to store million pages of paper documents into electronic form, OCR is the key tool. Otherwise, if those are entered by typing manually, the efficiency, effectiveness and correctness will drastically fall down. As a result, all the effort will go in vain.

In any language, there are two types of written document. One is hand-written document and the other is printed document. Most of the characters of Bangla hand-written words are touching which is the prime bottleneck of this kind of documents. Besides, the shapes of Bangla hand-written characters are extremely diversified. Doing proper segmentation of these scripts is really tough. Hence, creating an efficient and useful OCR system for Bangla handwritten scripts is really a great challenge for computer scientists. On the other hand, the printed Bangla documents are much simpler than that of the previous one; because, in printed scripts variations in style of Bangla characters are limited. The primary alphabet of Bangla script is quite large compared to the alphabet sets of English and other western languages. It comprises of 11 vowels, 39 consonants and 10 numerals. The total number of symbols is approximately 300. Besides this huge quantity of symbols, there are various types writing style of those. All these aspects have thrown a great challenge to the researchers in developing a comprehensive OCR for Bangla printed scripts. For both on-line and Off-line OCR, recognizing the Bangla scripts is really tough, some sophisticated research and development has been done on recognition of handwritten Bangla numerals, but ,cam, research works have been found on overall Bangla OCR.

1.4 Problem Definition

The majority of the data available today is either on paper or as archive or still photos. Text in images and documents contain meaningful and useful information that can be used to fully understand the contents. Text extracted from images play an important role in document analysis, vehicle plate detection, video content analysis, document retrieval, blind and visually impaired users etc. A document image contains various information such as texts, images and graphics i.e., line drawing and sketches. They are developed by scanning journals, historical document images, degraded document images, handwritten text images printed document images, multi-color book cover, newspaper etc. But important document can be destroyed or distorted in many ways. Problems also arise due to low resolution of the camera used. In old manuscript degradation of text may arise due to environmental effect of air, moisture, dust, rain, heat etc. These effect may

change the original document to a form where it is strongly needed to enhance the quality of the image of manuscript and then trying to recognize the character in the enhanced image. Besides the difference in boldness of the character, different font of a single character, its size and skewness makes it even harder to recognize. In case of multi-color document image, extraction of text is difficult and unreadable due to color mixing of foreground text with background. Generally Optical Character Recognition (OCR) system works on a screen background images. Indeed, even great quality optical character recognition system performs inadequately during the content extraction in the case of such issues. So to keep the vital data loss of these archives and images and furthermore to fabricate advanced libraries this expansive volume of data should be digitalized into images, upgrade the images if necessary and afterward content to be recognized and trained.

1.5 Background Study

In recent years, OCR has been an important application of pattern recognition as well as computer vision. Although the history of research on complete handwritten Bangla recognition is not outstretched, a few breakthroughs have already been found. Different authors proposed multifarious strategies for recognizing Bangla characters. A.F.R. Rahman et al. developed a multistage approach which includes Matra- upper part of the character, disjoint section of the character and vertical line for OCR of Bangla basic characters [3]. N. Das et al. proposed a system using Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) in the detection of both Bangla basic and compound characters [2]. S. Bhowmik et al. worked with MLP algorithm using HOG feature and achieved 87.35% accuracy but they approached to identify holistic words [4]. M. M. Alam et al. proposed a system for printed characters using Freeman chain code for representing a character and MLP is used to classify characters [1]. M. M. Islam et al. developed a system of Automatic Feature Extraction using XOR Operation for Bangla character recognition [5]. S. Afroge et al. came up with feature extraction scheme based on “Discrete Frechet Distance” and “Dynamic Time wrapping” [6]. In fact, classification of Bangla characters evolved over time to recognize characters [7, 8]

1.6 Objective of the Thesis

The fundamental goal of this theory is to get the editable text from picture and reports of handwritten manuscript. The objectives of the thesis is given bellow:

- Determine each letter in an of manuscript
- Remove noise and scale it.
- Recognition of each letter of the manuscript.

1.7 Decomposition of the Thesis

This thesis comprises of following substance:

Chapter 1- This chapter introduces basic Bangla characters, history of Bangla scripts, problems of recognizing handwritten Bangla character recognition and the background study of this research area of handwritten Bangla character recognition.

Chapter 2- Image, Digital image and its different types, Image processing, Image enhancement and different type of image enhancement operations, translation, rotation, scaling

Chapter 3- Preprocessing, Segmentation, Feature extraction, Recognition techniques are discussed.

Chapter 4– The proposed system is elaborately clarified with necessary figures and brief demonstration.

Chapter 5- Experimental Results and Performance analysis, Comparison of different feature extraction techniques are briefly described.

The **references** of the paper are listed in the last section of this book.

Conclusion

The essence of Bangla character recognition is depicted in this chapter. The research history of Bangla OCR clearly indicates that a stable Bangla OCR can be expected to find in near future.