5

On OCR of a Printed Indian Script

Bidyut B. Chaudhuri

5.1 Introduction

One of the earliest practical problems solved successfully by the pattern recognition community is the optical character recognition (OCR) whereby text in a document is automatically converted into electronic format [16]. The application potentials of OCR have been soundly established over the past thirty years and many commercial products are available in the market for business letter reading, table form processing, postal address reading, signature verification and reading aid for the blind. Excellent surveys on the OCR research are available in [12–14, 21].

For recognition tasks, we can distinguish between machine-printed and handwritten texts. The latter can be further subdivided into on-line or off-line handwritten text recognition. We are concerned here on the OCR of machine-printed Indian text only. The problem is very important since it is largely unsolved, although Indian languages have a following of more than one billion people and there are more than a dozen of major Indian scripts used in this country. Here the problem is challenging because Indian scripts consist of large collection of compound characters, and as described in section 5.2 next section, there may be more than one thousand different shapes to be recognized by the OCR system.

Till 1990, recognition studies on Indian scripts were concentrated on a small subset of manually segmented basic characters only [24–26]. The development of a complete OCR system from real documents with good accuracy was first reported for single font printed Bangla and Devanagari text in the middle of 1990s [5, 17–19]. The work has since been extended to other major Indian scripts like Oriya, Punjabi, Telugu and Tamil [1, 6–8, 11, 15, 17, 18, 22] with various degrees of success. Work on handwritten character recognition was also attempted [2, 3, 20].

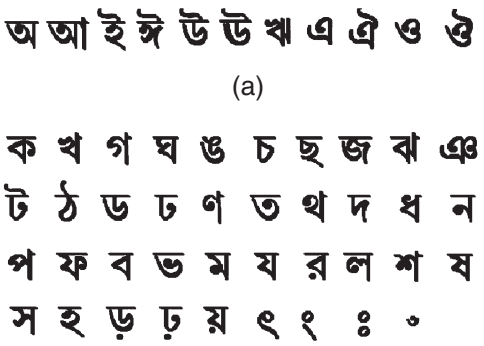

Fig. 5.1. Basic Bangla characters: (a) vowel, (b) consonant and (c) vowel allograph.

This chapter will describe briefly the problem of printed Indian script OCR and concentrate on the Bangla Script. The basic alphabet of Bangla script is shown in Figure 5.1. Bangla and Devanagari are the most popular scripts used for writing the languages of more than half of Indian subcontinent. The choice of Bangla is made by the fact that it is one of the most difficult Indian scripts from OCR point of view and it will shed light on different problems that may be encountered in Indian script OCR effort. Moreover, Devanagari script is quite similar, but somewhat simpler than Bangla. So, techniques of Bangla OCR could be used for Devanagari with little modification.

The rest of this chapter is organized as follow. The origin and properties of Indian Scripts are described in Section 5.2. Section 5.3 deals with document pre-processing approaches. Feature extraction and recognition of characters are described in Section 5.4. Section 5.5 treats the performance analysis of such an OCR system. Concluding remarks are given in Section 5.6.

5.2 Origin and Properties of Indian Scripts

Scholars agree that most alphabetic scripts are derived from the ancient Semitic alphabet that originated in the second millennium BC in and