Chapter 14
Devnagari Script Recognition: Techniques and Challenges

P. Mukherji
University of Pune, India

P.P. Rege
College of Engineering Pune, India

ABSTRACT

Devnagari script is the most widely used script in India and its Optical Character Recognition (OCR) poses many challenges. Handwritten script has many variations, and existing methods used are discussed. The authors have also collected a database on which the techniques are tested. The techniques are based on structural methods as opposite to statistical methods. There are some special properties of Devnagari script like the topline, curves, and various types of connections that have been exploited in the methods discussed in this chapter.

INTRODUCTION

In this chapter, techniques for Devnagari Handwritten Script Recognition are discussed. To accommodate the variations in handwritten characters Adaptive Zoning (AZ) method is proposed, where first a few features are examined and then, the region of interest of the character is isolated. Devnagari characters are analyzed on the basis of shape descriptions of provided by Devnagari script writers. A novel algorithm of segmenting a character in strokes and encoding the strokes into meaningful entities by Average Compressed Direction Coding (ACDC) algorithm is explained in Stroke Analysis method. Distribution of statistical and location features of strokes are compared with the stored model. Fuzzy location and stroke features are also used to include t he variations in handwriting. Inter-stroke properties of Devnagari characters are studied by generating Hierarchical Attributed Relational Graphs (HARG) method. There are two levels in this scheme in which first the sub-graphs are identified and then the graph is matched by converting it into a vectorial signature. Concept of segments and complement segments is introduced and their interconnections give the segment adjacency. The recognition accuracy achieved is comparable to recent published works.

DOI: 10.4018/978-1-61350-429-1.ch014

BACKGROUND

Optical Character Recognition (OCR) is the study of teaching machines to observe the environment and learn to read characters and make decisions. Character and pattern recognition are basic requirements in Artificial Intelligence. A character also comes in the general category of a pattern. In Jain A. K., Duin R. P. W. & Mao J. (2000), pattern is defined "as opposite to chaos; it is an entity and could be given a name".

OCR Basic Principles

Handwritten or typed data is converted to digital form either by scanning the writing on paper or by writing with a special pen on an electronic surface such as a digitizer combined with a liquid crystal display. The two approaches are distinguished as off-line and on-line OCR Plamondon R. & Srihari S. N. (2000), respectively.

Prior to feature extraction, preprocessing improves recognition efficiency. Preprocessing includes noise removal, machine and handwritten character segmentation, script identification, graphic and text segmentation and all such techniques that lead to improved recognition accuracy.

Feature extraction based methods work on extracting a set of invariant features from the test pattern and the classification is done in feature space.

Character classification can be achieved in two stages: coarse classification and fine classification. Coarse classification is accomplished by class set partitioning or dynamic character selection Duda R. O., Hart P. E. & Stork D.G. (2001). A tree classifier Gonzalez R. C. & Woods R. E.(2003) is used to selectively examine presence or absence of certain feature at each node thereby reducing the search.

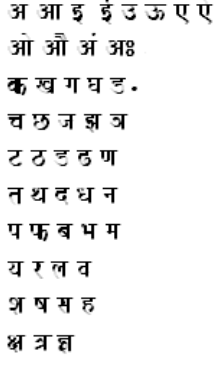
The Devnagari Script

Devnagari script is the most widely used script in India. Just as Kanji is used in Japanese and Chinese language, Devnagari is used in over forty languages including Sanskrit, Hindi, and Marathi etc.

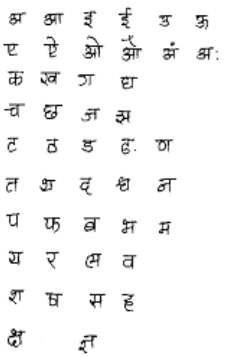
The basic character set of Devnagari script is of 48 characters and Shivaji 01 font is shown in Figure 1(a). The character set of Devnagari script with 45 characters is shown in Figure 1(b).

Every individual word has a horizontal header line or the 'shirorekha'. This line serves as a reference to divide the character into two distinct portions: Head and Body, if the top modifier is present. Devnagari word may be divided in three zones. Zone 1 is the region of top-modifier; Zone 2 is the body of the word and Zone 3 is the lower modifier region. Another feature is the inter-character gap in a word that facilitates character segmentation and isolation.

Figure 1. Devnagari character set



(a) Devnagari Character Set(Printed)



(b) Devnagari Character Set (Handwritten)