Chapter 5
Selection of an Optimal Set of Features for Bengali Character Recognition

Hasan Sarwar
United International University, Bangladesh

Saima Hossain
LEADS Corporation Limited, Bangladesh

Mizanur Rahman
Institute of Science and Technology (IST), Bangladesh

Sabrina Ahmed
Local Government Engineering Department (LGED), Bangladesh

Nasreen Akter
St. Francis Xavier University, Canada

Chowdhury Mofizur Rahman
United International University, Bangladesh

ABSTRACT

Feature extraction is an essential step of Optical Character Recognition. Accurate and distinguishable feature plays a significant role to leverage the performance of a classifier. The complexity level of feature identification algorithm differs for alphabet sets of different languages. Apart from generic algorithms to find features of different alphabet sets, these algorithms take care of individual characteristic common for a particular alphabet set. Dominant features of one alphabet set might completely differ from that of another set. Since there always remains the chance that inaccurate features may cause inefficient recognition, special attention should be given to identify the set of optimal features of a character set. Bengali characters also have some specific issues apart from the existing issues of other character sets. For example, there are about 300 basic, modified, and compound character shapes in the script, the characters in a word are topologically connected, and Bengali is an inflectional language. Literature survey shows that several authors have used different features and classification algorithms. The authors have extensively reviewed all these feature sets. In order to identify an optimal feature set, variability analysis has been proposed here. They focus on the specific peculiarities of Bengali alphabet sets, its different usage as vowel and consonant signs, compound, complex, and touching characters. The authors also took care to generate easily computable features that take less time for generation. However, more attention needs to be given in order to choose an efficient classifier.

DOI: 10.4018/978-1-4666-3970-6.ch005

1. INTRODUCTION

1.1. Background

The necessity to have a workable Bangla Optical Character Recognizer (OCR) is felt by all Bengali speaking people. Bangla (or Bengali), one of the more important Indo-Iranian languages, is the sixth-most popular in the world and spoken by a population that now exceeds 250 million. Geographical Bangla-speaking population percentages are as follows: Bangladesh (over 95%), and the Indian States of Andaman & Nicobar Islands (26%), Assam (28%), Tripura (67%), and West Bengal (85%). The global total includes those who are now in diaspora in Canada, Malawi, Nepal, Pakistan, Saudi Arabia, Singapore, United Arab Emirates, United Kingdom, and United States. The history of developing a complete Bangla Optical Character Recognizer (OCR) dates back to 1990s(Alam & Kashem, 2010; Chaudhuri & Pal, 1998; Mahmud et al., 2003; Omee et al., 2011). Recent trend of digitizing knowledge repositories, implementation of E-Governance requires easy entry of already printed data into computer. A lot of research effort has already been committed to solve many individual intricate issues. A detailed summary is found here (Hossain, Akter, Sarwar & Rahman, 2010). Research toward a commercially viable Basic OCR still deserves considerable amount of effort. A survey of the existing literature exhibits that most of the authors have tried to develop an OCR considering all the steps of a pattern recognition system. For example, all published literature have considered the tasks of preprocessing, noise elimination, skew detection, segmentation, feature selection and extraction, classification and post processing. A complete in-depth focus on any particular aspect, involving the inherent nature of Bangla Language, is still missing in the whole spectrum of research works in this particular area. Here, we have tried to focus on the feature identification part of the whole issue. In our view, this process deserves special attention and care to be paid in order to build a successful OCR. In our knowledge, this is the first attempt so far taken in the domain of Bangla OCR development.

1.2. A Brief on Bangla Alphabet

The basic Bengali character set comprises of 11 vowels, 39 consonants, and 10 numerals. There are also compound characters being combination of consonant with consonant as well as consonant with vowel. A vowel following a consonant sometimes takes a modified shape and is called a vowel modifier. Similarly there are consonants take the shape of a modifier when comes with another consonant. In general, there are 300 basic, modified and compound character shapes. These characters in a word are topologically connected. Bangla is known as an inflectional language. Some of the characters are shown in Box 1.

Box 1.

Vowels অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

Vowel Modifiers া ি ী ু ূ ৃ ে ৈ ো ৌ

Consonants ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ ব ভ ম য র ল শ ষ স ড় ঢ় য় ৎ ং ঃ ঁ

Numerals ০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯

Some Compound Characters ক্ষ ঙ্ক ঙ্গ ক্ত ঞ্চ ঙ্ক্ষ ঞ্ছ ঞ্জ ন্ধ্র ষ্ঠ