Optical Character Recognition for Brahmi Script Using Geometric Method

Neha Gautam and Soo See Chai
Faculty of Computer Science and Information Technology, University Malaysia Sarawak.
nehagautam1208@gmail.com

***Abstract*—Optical character recognition (OCR) system has been widely used for conversion of images of typed, handwritten or printed text into machine-encoded text (digital character). Previous researches on character recognition of South Asian scripts focus on modern scripts such as Sanskrit, Hindi, Tamil, Malayalam, and Sinhala etc. but little work is traceable to Brahmi script which is referred to as the origin of many scripts in south Asian. This study proposes a method for recognition of both handwritten and printed Brahmi characters which involve preprocessing, segmentation, feature extraction, and classification of Brahmi script characters. The geometric method was used for feature extraction into six different entities, followed by a newly developed classification rules to recognize the Brahmi characters based on the features. The method obtains accuracy of 91.69% and 89.55% for handwritten vowels and consonants character respectively and 93.30% and 94.90% for printed vowel and consonants character respectively.**

***Index Terms*—OCR; Brahmi Script; Geometric Features; Zone Method; Asian Scripts.**

I. Introduction

OCR has provided an efficient method to handle Character recognition [1]. OCR is gaining an increasing importance because of the demand for creating a paperless world and digitization [2]. OCR process belongs to the family of techniques used for performing automatic identification and Automatic script recognition [2]. OCR provides the solution for automatically processing large volumes of data.

Despite its usefulness, OCR is not been successfully adapted to recognizing printed or handwritten document or images of varying scripts [3, 4]. The research focus in the past decades has been to develop more algorithm using this technique for script identification [4].

A study by Trautmann and Thomas [5] identified 198 different modern scripts which originated from Brahmi script in the South and Central Asia. Scripts such as Devanagari, Bangla, Gurmukhi, Gujarati, Oriya, Kannada, Telugu, Tamil, Malayalam, and Urdu are referred to as modern script according to Pal, Jayadevan, and Sharma [6]. Winskel and Padakannaya [7] noted that there are similarities in the structural-features (straight, slant lines, and curves) of the modern-Asian scripts (Hindi, Sanskrit, Sinhala) and Brahmi script.

There is need to develop an efficient method to automatically extract features and recognize the characters of Brahmi script. Existing algorithms for feature extraction uses geometric properties and invariant techniques based on shapes [8]. Geometric features are features of objects constructed by a set of geometric elements like points, lines, curves, surfaces, corner, and edge, etc. which can be detected by feature extraction methods [9].

In this study, the geometric method was used for feature extraction and followed by a newly developed classification rules to recognize the Brahmi characters based on the features. According to this approach, Brahmi characters was identified with good accuracy. Brahmi script character recognition is important in the field of archaeology and epigraphy [10]. It can help to find the relationship between Brahmi script and another script of South Asia [11, 12].

II. Literature Survey

A. Brahmi script recognition

Siromoney et al. [13] used the coded run method for the recognition of machine-printed characters of the Brahmi alphabets. Each Brahmi character is changed manually into a rectangular binary array, this method can be applied to any script. In 2006, Devi suggested two methods for preprocessing part of Brahmi character recognition: thinning and thresholding method [14, 15]. The analysis of results was done by preprocessing pixel-level technique for Brahmi script in OCR system [14]. It involves a cascaded approach in which various thinning and thresholding algorithms are applied on the input image [15]. Gautam, Sharma, and Hazrati [16] obtained accuracy of 88.83 % using zone method for the feature extraction and template matching method (lower and upper approach) for the classification of handwritten Brahmi character recognition. However, non-connected characters could not be recognized using this method [16].

B. Geometric method for feature extraction

Gaurav and Ramesh [17] applied the geometric method for feature extraction of English character recognition by using starters, intersections, minor starters etc. The method was tested after training a Neural Network with a database of 650 images. In 2013, Dongre and Mankar used geometrical feature extraction (such as Horizontal lines, vertical lines, etc.) to recognize the Devanagari characters. The accuracy obtained in the study can be improved upon, by using ANN and SVM classier [18]. According to Akram, Bashir, Tariq, and Khan [19], the feature vector for each English character contains different features (number of endings, corners, and bifurcations) and the characters can be recognized via simple rules which are based on extracted features. However, the study did not clearly discuss how to recognize characters with the same type of features. Dongre and Mankar [20] considered geometric features (horizontal lines, vertical lines, etc.) for the recognition of isolated Devanagari characters with the accuracy of 93.17%. The major problem in the study by [20] is the recognition of conjuncts and compound characters which denote connectivity with the vowels and consonants, making them conjuncts and compound