A Brief Study of Feature Extraction and Classification Methods Used for Character Recognition of Brahmi Northern Indian Scripts

Rohit Sachdeva, Asstt. Prof., Computer Science Department, Multani Mal Modi College, Patiala
Dharam Veer Sharma, Asstt. Prof., Department of Computer Science, Punjabi University, Patiala

ABSTRACT

According to the 8th schedule of Indian constitution, there are 22 official languages and 122 regional languages prevalent in India. In the last few decades, the recognition of these scripts has been prominent area of research. Among these scripts most of the recognition research work has been done for Bangla, Devanagari, Gujarati, Gurmukhi and Telugu scripts etc. Commercial OCRs were available for various scripts like Latin, Japanese, Chinese, Roman, Arabic scripts. OCR systems for few Indian scripts are available and others are in the stage of development for preserving manuscripts and ancient literatures written in different Indian scripts and making digital libraries for the documents. Further, overall accuracy of the recognition, feature extraction and classification are crucial phases. This paper attempts to give a brief summary of various feature extraction and classification methods used for recognition process of Brahmi Northern Indian scripts by the researchers in last few decades.

Keywords
Optical Character Recognition (OCR), Brahmi Northern Indian Script, Character Recognition, Feature Extraction, Classification.

INTRODUCTION

The first idea of the concept of the OCR was given by Tausheck [1] and Handel [2]. Early version of OCR research was to recognize characters from images of machine printed text with single size and single font. Optical Character Recognition (OCR) is an approach that converts printed, typewritten or handwritten into editable form, which can be further used as per necessity. OCRs are divided into two sub categories: type written or printed text and Hand written text. Handwritten Text OCR can further divided into two sub categories: Online Recognition and Offline Recognition. Later is a process which means that after the completion of writing or printing, the recognition starts. During the last three decades, character recognition research has been prominent area of research field.

India is a multi-lingual or multi script country. According to the 8th schedule of Indian constitution there are 22 official languages and 122 regional languages. Most of the Indian scripts have their origin from Brahmi script, through with certain alterations. Brahmi is the ancestor of hundreds of languages predominantly used in the Indian sub-continent as well as in South-East and East Asia. In India, Brahmi script divided zonally into two sub scripts: Northern and Southern Script. Bengala, Devanagari, Gujarati, Gurmukhi and Oriya are Northern Scripts and Kannada, Malayalam, Tamil and Telugu are Southern Scripts.

Development of OCR system for these Indic Scripts has several application areas such as preserving manuscripts and ancient literatures written in different Indian scripts and making digital libraries. OCR system comprises the following steps:

1. Image Digitation
2. Preprocessing
3. Feature Extraction
4. Classification
5. Post-Processing

1. Image Digitation:
Image of the source document is firstly scanned and stored in some image file in the form of bitmaps. This is also called digitization. With the aid of scanning method, digital image of the source document is captured.

2. Pre-processing:
In pre-processing, some pre-processing of the image, having text or data which has to be recognized, is required for improving the recognition accuracy. The pre-processing activities may include the following: noise removal, document level skew detection and correction, binarization of the digitized image, size normalization and segmentation at all the levels i.e. line, word and character.
3. **Feature Extraction:**
The main objective of feature extraction is to capture the vital characteristics of the symbols. It is the crucial and most vital stage of recognition process. Under this step, the features are extracted from segmented symbols. So, to attain the high recognition percentage, selection of feature extraction method becomes vital factor. The extracted features may be structural, statistical or moments based. Some feature extraction methods are Contour Profile, Deformable Templates, Moments calculation (Ex-Geometrical, Hu-moments, Zernike), Projection Histogram, Template matching, Zoning etc.

4. **Classification:**
In the classification step, the features extracted from previous steps i.e. feature extraction step are used to recognize the text segment according to the preset rules. Classification is usually done by comparing the feature vectors corresponding to the input character with the representative(s) of each character class, using a distance metric. It is the procedure of assigning the detected data to their corresponding class with respect to groups with homogeneous characteristics, with the aim of cultivated numerous objects from each other within the image. It is carried out on the basis of stored features in the feature database, such as global and structural features etc. On the basis of decision rule, classification divides the feature space into several classes. Various classification procedures used in earlier developed Optical character recognition systems are Bayesian Classification, Decision Tree Classification, K-Nearest Neighbors, Neural Network, and Support Vector Machine.

5. **Post-Processing:**
Post-processing step involves grouping of symbols. The process of performing the association of symbols into strings is referred to as grouping.

**FEATURE EXTRACTION AND CLASSIFICATION METHODS USED IN CHARACTER RECOGNITION FOR BRAHMI NORTHERN INDIAN SCRIPTS**

**Bangla**
Chaudhuri et al. [3] represented a bilingual OCR system which is used for recognize Bangla and Devanagari script. A headline deletion process was used for character segmentation. For easier recognition, a text line was divided into three sub zones. For the recognition of basic and modified characters, structural feature and binary tree classifier was used and for the recognition of compound character, a hybrid method combining structural and run based template features was used.

For Bangla script Jalal et al. made very prominent effort. They presented the system which used Bounded rectangle calculation, Chain code generation; Slope distribution generation features extraction methods along with neural network classifier. Authors claimed that their system attained 96% of accuracy.
U. Bhattacharya et al. [5] have given a system for recognition of Handwritten Bangla Characters. They used local chain-code histograms for obtaining features of input character and MLP classifier. They claimed that their system achieved 92.14% accuracy on testing sets and 94.65% on training sets.

Devanagari

M. K. Sinha et al. [6] proposed a template based OCR system for handwritten Devanagari documents. In term of primitives and relationships, the system stores structural descriptors for each symbol of the script.

Veena Bansal et al. [7] presented hybrid classifier-based complete OCR for printed Hindi text written in Devanagari script. This system also supports the touching characters and compound characters in noisy environment. For the character segmentation, a projection profile technique was used by them. The system used following multiple features extraction method such as coverage of the region of the core strip, Horizontal zero crossings, Moments, Number of positions of the vertex points, Structural descriptors of the characters for classification, Vertical bar feature along with hybrid classifiers. At the character level overall accuracy attained by system was 93%.

Reena Bajaj et al. [8] have proposed system, to recognize handwritten numeral of Devanagari script. They suggested a method for recognition of handwritten Devnagari numerals using density, moment of curve and descriptive component feature with MLP classifiers.

Gujrati

Antani et al. [9] have proposed the classification of a subset of printed or digitized Gujarati characters. Euclidean Minimum Distance, Hamming Distance classifier and K-Nearest Neighbour classifier have been used for classification with template matching. But a very low recognition rate of 67 percent is reported.

Yajnik et al. [10] have developed a system for classification of Sets of printed Gujarati characters and modifiers using ANN architectures by considering linear activation functions in the output layer. Printed Gujarati text features were extracted in terms of wavelet coefficients. They have used two Multi-Layer Perceptron (MLP) networks which are used for classification of alphabets in middle zone and lower zone separately. These networks achieve 94.46 percent accuracy for middle zone and 96.32 percent accuracy for lower zone alphabets and modifiers.

Prachi et al. [11] proposed a Gujrati OCR system for the recognition of basic characters in printed Gujarati text. Principal Component Analysis (PCA) was used to extract the features of printed Guajarati characters. For the classification of characters based on features Hopfield Neural classifier had been used by them. The system attained the 93.25% accuracy.

Gurumukhi

G. S. Lehal et al. [12] proposed a OCR system for Gurumukhi script. They used Local features such as branches, concave/convex parts, joints, number of endpoints etc. and Global features such as connectivity, number of holes and projection profiles etc. along with hybrid classification technique such as binary decision tree and Nearest Neighbour classifiers. They achieved a recognition rate of 91.6%.

Dharam Veer Sharma et al. [13] used zoning feature with hybrid classification technique using K-Nearest Neighbour and Support Vector Machine classifier. But a very low recognition rate of 67 percent is reported.

Geeta et al. [14] proposed an OCR system for Gurumukhi numerals. They used Zone Distance features and Support Vector Machine classifier. They stated that their system attained 99.73% accuracy.

Oriya

B Chaudhuri et al. [15] presented a model for Oriya script OCR. They used Directional as well as Global Features and classified them using Decision tree classifier. They attained 96.03% accuracy at character level.

For off-line unconstrained Oriya handwritten numerals a system was proposed by Roy et al. [16]. They used histograms of direction chain code of the contour points of the numerals as features and a neural network based classifier. They attained 94.81 % accuracy.

CONCLUSION WITH COMPARISON TABLE

A brief study of feature extraction and classification methods used for character recognition of Brahmi Northern Indian Scripts shown in a tabular form is given below.
This study will definitely help developers and research scholars in the concerned area. OCR for these scripts that works under all possible circumstances and gives highly precise results, is the area which needs to be explored more to get more precise results. By using the hybrid methods, higher recognition rate could be attained.

**REFERENCES**


