Character Recognition of Gujarati and Devanagari Script : A Review

S. S. Magare
Dept. of CS-IT
Dr. B.A.M. University
Aurangabad (M.S)

Y. K. Gedam
Dept. of CS-IT
Dr. B.A.M. University
Aurangabad (M.S)

D. S. Randhave
Dept. of CS-IT
Dr. B.A.M. University
Aurangabad (M.S)

Prof. R. R. Deshmukh
Dept. of CS-IT
Dr. B.A.M. University
Aurangabad (M.S)

Abstract

In this paper, we describe the different techniques of character recognition for Gujarati and Devanagari script. Character recognition is usually referred to as OCR. Review of this paper will provide a way for researcher to develop a tool for Gujarati and Devanagari script recognition. This paper describes basics of character recognition, its type, challenges associated with it and the special properties of Gujarati and Devanagari script.

Keywords

Offline character recognition, Online Character recognition, Handwritten character recognition, Printed character recognition, feature extraction techniques, classification.

1. Introduction

Optical character recognition is the process of recognizing optically scanned characters. Character recognition has two types: Offline and Online .One of the challenging problem in pattern recognition is Offline character recognition. Offline character recognition takes scanned image of required document paper. Scanned image can be in color form. For the reorganization process this image is converted to grayscale and then Binarization is applied on grayscale image. So that image can contain information only in 0 or 1.Offline character recognition can be done in two ways: Handwritten and Printed.

Handwritten character recognition is abbreviated as HCR; handwritten characters have number of variations as different people has different writing styles. HCR can recognize offline character and online characters. Offline HCR takes input from scanned image of paper document and Online HCR takes input from digital pen. There are many handwritten historical documents exist in electronic form, HCR is used to recognize such documents.

1.1 Gujarati Script

Gujarati script is derived during 16th century from Devanagari script and it is modern language of India. Main difference between Gujarati and Devanagari script is the lack of horizontal line at header of character in Gujarati script and small modification in the characters. Until the 19th century Gujarati script was only used for accounting and writing letters, because of this Gujarati script was also known as Banker's, Merchant's and traders script. Recognition of Gujarati script

1.2 Devanagari Script

Devanagari Script consists of 14 vowels and 34 consonants. Devanagari script is base for writing 28 languages such as Marathi, Hindi, Sanskrit, Kashmiri, Bhojpuri and many more. Devanagari script is formerly used to write Guajarati. It is written from left to right. Devanagari script consists of horizontal line at the header of character called Shirorekha. Devanagari is most used and adopted writing system. Character recognition of Devanagari Script is somewhat challenging due to curve involved in most characters.

Figure 1. Consonants of Gujarati Script

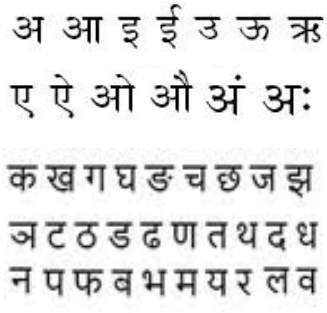

Figure 2. Vowels and Consonants of Devanagari Script

2. HCR system

The block diagram of HCR system is shown in fig.1. It consists of Preprocessing, Segmentation, feature extraction, classification and recognition.

2.1 Preprocessing

Preprocessing technique is used to do improvement of image data that enhances some image features required for processing and suppresses unwanted noise and distortion from image data and aims to correct degradation in an image

2.1.1 Binarization

Binarization is the process of converting grayscale image in to binary (Black and White) image, so that image data will only contain 0 and 1. Binarization technique is usually used for separating foreground from background using required level of thresholding.

2.1.2 Noise Removal

Digital image consist of variety of noises. These noises are required to be removed from an image for better processing. Morphological operation, Median filter and Weiner filter is used to remove noise from an image. Median filter reduces blurring of edges.

2.1.3 Thinning and Filling

Smoothing implies both Filling and Thinning. Thinning reduces width of character while Filling eliminates gap, small breaks and holes in digitized character.

2.1.4 Normalization

To obtain characters of uniform size, rotation and slant Normalization is applied on image. To improve the accuracy of character recognition Normalization reduces shape variation.

2.1.5 Skew detection and correction

During the digitization of document page it is often that image is not align correctly or it may be happen by human while writing document. To make in correctly align Skew detection and correction technique is used.

Skew detection technique can be classified in to groups: Analysis of Projection profile, Hough transform, clustering, connected component and correlation between line techniques.

2.2 Segmentation

Segmentation of an image is the process of subdividing image into number of parts. Segmentation takes the form as Paragraph Segmentation, Line Segmentation, Word Segmentation and Character Segmentation.

Paragraph wise segmentation divides the document into paragraph. Line wise segmentation divides paragraph into line. Line wise segmentation can use a horizontal projection profile based techniques Word wise segmentation divides line into word. Finally, Character wise segmentation divides words into characters.

Chain code histogram can be used for each segment. Horizontal projection file method is used for segmentation.

2.3 Feature Extraction

Feature extraction technique is aims to extract the essential and important features and characteristic of the given image. In Pattern recognition this is one of the difficult stages to implement. Selection of right feature extraction technique leads to achieving high performance for recognition.

Feature extraction technique is divided into three groups: Distribution of points, Transformation & series expansion and Structural analysis. Structural analysis extracts the feature which represents geometric and topological structure of character. Structural analysis gives feature with high tolerance of noise and style variation. Commonly used features are intersection between lines and loops.

Table 1. FE method for various image representation forms

Feature Extraction Method	Gray scale sub-image	Binary Image		Vector (skelton)
		Solid character	Outlet Contour	
Template matching	Yes	Yes	No	Yes
Deformable templates	Yes	No	No	Yes
Graph description	No	No	No	Yes
Unitary transform	Yes	Yes	No	No
Discrete features	No	No	No	Yes
Zoning	Yes	Yes	Yes	Yes
Fourier descriptors	No	No	Yes	Yes
Geometric moments	Yes	Yes	No	No
Zernike moments	Yes	Yes	No	No
Projection histogram	No	Yes	No	No
Contour profile	No	No	Yes	No
Spline curve	No	No	Yes	No

2.4 Classification

After selection of the features next step is to classify them according to its properties. Training and testing is done at the classification phase. Number of classifier can be used to train the character. K-NN method is mostly used at classification stage.

2.4.1 Neural network

Neural network is one of the well known classifier used for character recognition system. Neural network have advantage of their adaptive nature. Feed forward NN and Back propagation NN is used for character recognition.

2.4.2 SVM

Support vector machine construct the hyper-planes in high or infinite dimensional space. SVM is based upon statistical learning theory. The SVM was defined for the two class problem and it looked optimal hyper-plane, which maximized the distance, margin, between the nearest examples of both classes.

3. Techniques used for scripts

3.1 Gujarati Script

In paper [1] presented Zone identification technique. Zone identification technique identifies three zones from Gujarati characters i.e. Base character zone, Upper modifier zone and Lower Modifier zone and Lower zone. They have found that several characters are discriminated by specific modifier, which exist in upper and lower zone. Therefore they have used Zone identification technique.

Thinning & skew correction is used for preprocessing and use Multi Layered Feed Forward Network for classifying digits.

3.2 Devanagari Script

In Paper [3] has described various Feature Extraction Method, such as Template matching, Deformable templates, Graph description, Discrete features, Zoning and Fourier descriptor. They found that Real- Valued feature vectors are ideal for statistical Classifier.

Chain coding used to extract chain code features at the feature extraction stage and use Combined MLP and Minimum Edit Distance Classifier for classification [4].

In paper [5], Median and Wiener filter for de-noising. They have used Structural segmentation algorithm for segmentation purpose and for feature extraction they have used Zone based approach.

Encode binary variation method for extracting the features. For classification purpose use SVM Comparison Techniques [6].

Segmentation based on character height and width [6]. At the classification process they have used MLP learning algorithm for two hidden layers with back propagation for character identification.

4. Dataset

There is no standard dataset available for handwritten characters. Researcher has to develop own character dataset collected from minimum 10-15 people. For better result and accuracy collect dataset from large number of people, as different people has different writing styles, it will include variation in the character which will be useful while training and testing phase for character recognition.

5. Acknowledgement

The authors would like to thank the University Authorities for providing the infrastructure to carry out the research. This work is supported by University Grants Commission.

6. References

[1] J. Dholakia, A. Negi, S. Rama Mohan. "Zone Identification in Printed Gujarati Text", *ICDAR*, Vol. 1, pp. 272-276, 2005.

[2] A. Desai, "Gujarati Handwritten Numeral Optical Character Recognition through Neural Network", *Pattern Recognition*, Vol. 43, pp. 2582-2589, 2010.

[3] Olivind Due Tier, Anil K Jain, Torfin Tax, "Feature Extraction Method For Character Recognition: A Survey", *Pattern Recognition* Vol. 29, No. 4, pp. 641-662, 1996.

[4] S. Arora, D.Bhattacharjee, M. Nasipuri, D. K. Basu & M. Kundu, "Recognition of Non-Compound Handwritten Devanagari Characters using a Combination of MLP and Minimum Edit Distance", *International Journal of Computer Science and Security (IJCSS)*, Vol. 4, Issue 1.

[5] Veena Bansal and R.M.K. Sinha, "Segmentation of touching and fused Devanagari character", *Pattern Recognition*, Vol. 35, Issue 4, pp. 875-893, 2002.

[6] U. Garain, B.B. Chaudhari, " Touching Characters in Printed Devanagari and Bangla Scripts Using Fuzzy Multifactorial Analysis", *IEEE Transaction on* Vol. 32, Issue 4, pp. 449-459, 2002

[7] Stuti Asthana, Farha Haneef and RakeshK Bhajade, "Handwritten Multiscript Numeral Recognition using Artificial Neural Nerworks", International Journal of Software Computing and Engineering ISSN: 2231-2307, Vol.1,Issue-1,March-2011.