Devanagari Letters Segmentation and Recognition System: A Brief Review

Shalini Dogra1 Amit Sehgal2
1Student 2Head of Dept.
1,2Department of Computer Science & Engineering
1,2SSGI, Yamunanagar (Haryana), India

Abstract—Character Recognition has been a significant area of study in Artificial Intelligence. The concept of Character Recognition became famous in recent times due to its many applications such as in printed postal addressing, filling of a variety of forms, multiple choice questions in certain examination and so on. OCR suggests to the technique of photo scanning of the text character-by-character, analysis of the scanned picture and then changing of the character image into information that a machine can easily read. e.g. organizations and libraries taking physical replicas of books, magazines, or other old printed material and utilizing OCR to place them into computers. Segmentation is the crucial and most complex part of OCR process, and it gets more difficult with handwritten text because of different writing styles and fonts used. Devanagari character recognition is more complex as it is having various loops, conjuncts, upper and lower modifiers and the number of disconnected and multistroke characters. This paper presents a comparative analysis of various character recognition techniques in terms of their accuracies and also their features, classifier and findings as well.

Keywords: Handwritten Character Recognition (HCR), Optical Character Recognition (OCR), Artificial Neural Network (ANN), Generalized Hidden Markov Model (GHMM), Generalized Regression Neural Networks (GRNN)

I. INTRODUCTION

A. Handwritten Character Recognition

Machine simulation of human reading has become a topic of serious research from the time when the introduction of digital computers. The main motivation for such an effort was not only the challenges in simulating human reading but also the possibility of efficient applications in which the information present on paper documents has to be transferred into machine-readable arrangement. Automatic recognition of printed as well as handwritten information present on documents like cheques, envelopes, forms, and other manuscripts has a range of practical and commercial applications in banks, post offices, libraries, and publishing houses. [4]

1) Challenges of Handwritten Character Recognition

As compared with printed character recognition, handwritten character recognition (HCR) is still a challenging task due to the following factors:

- individual styles of inscription,
- speed of writing,
- size of letters,
- physical and mental state of the writer,
- overlap of letters etc. [6]

2) Applications of Character Recognition System

Some of the vital applications of offline handwritten character recognition are listed in the following section:

a) Recognition of Ancient Document

The historical documents are currently being digitalized for avoidance purpose and to make them available universal through large on-line digital libraries.

b) Cheque Reading

Offline handwritten Character recognition is basically used for cheque evaluation in banks. Cheque reading is very important industrial application of offline handwritten character recognition. Handwritten character recognition plays very essential role in banks for signature authentication and for recognition of amount filled by user.

c) Postcode Recognition

Handwritten character recognition system can be used for evaluation the handwritten postal address on top of letters. Offline handwritten character recognition system used for recognition handwritten digits of postcode. HCR can be interpret this code and can sort mail without human intervention.

d) Form Processing

HCR can also be used for form processing. Forms are normally used for collecting public information. Responses of public information can be handwritten in the space given.

e) Signature Verification

HCR can also be used for identify the person by signature verification. Signature identification is the specific pitch of handwritten identification in which the author is verified by some specific handwritten text. [9]

3) Classification Handwritten character recognition

Handwritten character recognition is divided into two parts i.e. offline handwritten character identification and online handwritten character detection.

![Fig. 1: Handwritten character recognition](image-url)

a) Offline Handwritten Character Recognition

In this category of character recognition, the typed/handwritten character are scanned and then changed in to digital form. Offline character recognition is more demanding and difficult job as it does not have the benefit of recognizing direction of movements which writing the text.

b) Online Handwritten Character Recognition

In this category of online handwritten character recognition, inscription and recognition are done at the same time.
case, user will write character on any sensory part where sensor will pick up the marker action and after that it recognizes character on the basis of direction of the motion. Online character recognition is a lot easier than offline character recognition because there is timing information is accessible. This method is commonly available on touchpad; touch screen cell phones etc. [17]

B. Properties of Devanagari Script

The main feature of Devanagari Script is that it is written from left side to right side. In Devanagari Script, all characters/letters are connected together by a straight line at the top called as 'Shirekha' or 'Header Line' to form a single word. The term Devanagari is derived on or after the Sanskrit words "Deva + Nagari" i.e. (God + City); collectively they means, the script of the “City of the Gods”. Devanagari is the mostly popular language in India among all languages. Hindi, the popular language in India, is represented in Devanagari. Marathi, Nepali and Sanskrit are few more languages which are also represented in Devanagari.[4]

Devanagari has 13 vowels along with 14 modifiers of vowels and of “rakar,” as shown in symbols. [4] Vowels can be written as separate letters or with the help of some additional marks which can be printed before, after, above or below the consonants. When any Vowel follows a Consonant, then we can say that, it is written in his corresponding maatraa form. When vowels are written in this way, they are called as modifiers and the newly developed characters so formed are called conjuncts. Sometimes to develop a new shape or new character, two or more consonants are combined together. These new shaped group of characters are termed as compound characters. Devanagari Numerals are from 0 to 9. [10]

Another distinctive feature of Devanagari is the presence of a horizontal line on the top of all typeset. This stripe is known as header line or “shirorekha” (see Fig. 1.4). The words can typically be divided into three strips: top, core, and bottom, as shown in Fig.1.4. The header line separates the top and core strips and a virtual base line separates the core and lower strips. The top strip in general contains the top modifiers, and bottom strip includes lower modifiers. When two or more characters appear side by side to form a word in Devanagari, the header lines touch and generate a bigger header line. [4]

II. STAGES IN CHARACTER RECOGNITION

The process of changing of scanned image into a text document consist the following steps shown in the Fig.5

A. OCR System

In the design of OCR system following steps has been followed:

1) Preprocessing
2) Segmentation
3) Feature Extraction
4) Classification [19]

1) Preprocessing

The digital image obtained from scanning may include some amount of noise depending upon the quality of scanner. The removing of this noise from captured picture is called preprocessing. Noise might be skewed or broken character. The major benefit of pre-processing a handwritten character is to arrange the information so as to make the recognition simpler. Along with the noise reduction, normalization and segmentation of picture is also done to make the additional recognition process effortless. In normalization, resizing of characters is prepared. Normalization is done for stroke width, slant, slope, height of the characters. The normalization process will reduce each character image to one consisting of vertical letters of standardized height and made up of one pixel-wide strokes. [18]

2) Segmentation

Segmentation stage includes word segmentation, character segmentation as well as line segmentation. Modes for character segmentations are based on

- White space and pitch
- Projection analysis and
- connected component labelling [15]

3) Feature Extraction

Next step is extracting features from the segmented characters and to distinguish it from other character by comparing it with already stored patterns of all the characters in the library. There are various features used for character recognition. Key features may include height, width, density, loops, lines, stems and other character traits. Chain code histogram feature is used for recognition that is extracted by chain coding the contour points of the scaled character bitmapped image. View based features is also used that is extracted from scaled, thinned one pixel wide skeleton of character image. Shadow features are extracted from scaled character image. [17]
4) **Classification**

It is the decision making phase of a handwritten character recognition system.[17] Based on the extracted features, Classification is carried out.[19] The feature vector obtained from previous stages is assigned a class label and recognized using unsupervised and supervised techniques. The data set is separated into training set and test set for each character.[15] Classifier takes characteristics as input and gives the corresponding character class as output. Irrelevant features may be removed for better classification accuracy.[17]. Character classifier can be Bayes classifier, Nearest neighbour classifier, Radial basis utility, Support vector mechanism, Linear discriminant functions and Neural networks with or without back propagation.[15]

### III. LITERATURE SURVEY

The several approaches available for implementing the handwritten character recognition are described as follows:

#### A. Fuzzy Model

An approach to implement the Handwritten Character Recognition is provided by M. Hanmandlu, O.V. Ramana Murthy and Vamsi Krishna Madasini in 2007. This paper presents the recognition of handwritten Hindi Characters based on the modified exponential membership function integral to the fuzzy sets derived from features consisting of normalized distances obtained by the Box approach. The exponential membership function is modified by two structural constraints that are estimated by optimizing an objective function that includes the entropy and fault function. A Reuse Policy that offers direction from the past policies is utilized to get better the reinforcement learning. This relies on the past errors exploiting the precedent policies. The Reuse Policy improves the speed of convergence of the learning process over the approaches that learn without reprocess and combined with the use of the reinforcement learning, there is a 25-fold improvement in training. Experimentation is carried out on a database of 4750 samples. The overall recognition rate is found to be 90.65%. [1]

#### B. Artificial Neural Network

An approach to implement the Handwritten Character Recognition is provided by Raghuraj Singh, C. S. Yadav, Prabhat Verma and Vibhash Yadav in 2010. In this paper the authors proposed a technique for OCR System for different five fonts and sizes of printed Devnagari script using Artificial Neural Network. The recognition rate of the projected OCR system with the image manuscript of Devnagari Script has been found to be quite high. [3]

#### C. Bayesian Discriminator Function

An efficient character recognition system is proposed using LDA Analysis followed by a Bayesian discriminator function based on the Mahalanobis distance by Sajjad Pourmohammad, Reza Soosahabi and Anthony S. Maida in 2013. In this work, since LDA is tailored for Gaussian distributed data and the samples dimensionality is high, a couple of pre-processing steps have been applied to condense dimensionality and cluster the data into semi-Gaussian subclasses. In the initial step, affine transformations are applied to the training samples in order to formulate the scheme Robust against distortion. Scaling and Rotation are among those popular misrepresentations which have been considered in this work. Inactive pixels are cut off using a simple algorithm in the next Step. After that, principal component analysis (PCA) and k-means clustering are applied. The outcomes from pre-processing showed a great potential in dimensionality reduction using transformations that can maintain useful information. Numerical results on the MNIST dataset arrives at 3% error rate which is lower than the other linear approaches. The proposed linear techniques are discussed in a manner with the aim of make it easier to have a much clearer understanding of the technique and why it works compared to the other classification methods. [8]

#### D. Cluster Detection Technique

A new strategy for the segmentation of conjuncts, and overlapping characters in Devanagari script on Hindi language is proposed by Binny Thakral and Manoj Kumar in 2014. The proposed algorithm is focused around Cluster Detection technique and gives 95% correctness for segmenting touching, conjunct characters and 88% effectiveness for overlapping characters. [11]

#### E. Khmer Character Recognition

An approach to implement Handwritten Character Recognition is provided by Hann Meng and Daniel Moraru in 2014 is called Khmer Character Recognition (KCR) system implemented in Matlab environment using artificial neural networks. In this paper authors proposed, the KCR system described the utilization of integrated self-organization map (SOM) network and multilayer perceptron (MLP) network with back propagation learning algorithm for Khmer character recognition problem. [12]

#### F. Gradient Local Auto-Correlation

An approach to implement Handwritten Character Recognition is provided by Mahesh Jangid and Sumit Srivastava in 2014. This manuscript is focus on the utilization of object detection algorithm GLAC (Gradient Local Auto-Correlation) for the handwritten character recognition (HCR) crisis. HOG and SIFT are already used in this (HCR) field except GLAC which produced good results than HOG and SIFT for object detection problem like individual in images, pedestrian detection and picture patch matching. This paper utilized GLAC algorithm to recognize the handwritten Devanagari characters. GLAC functional on two handwritten Devanagari databases, ISIDCHAR and V2DMDCCHAR. The images of files are too normalized with and with no preserving aspect ratio. Using GLAC method and SVM classifier, the best results obtained on ISIDCHAR and V2DMDCCHAR are 93.21 %, 95.21 % respectively that justified the utilization of GLAC algorithm for character recognition problem. [13]

#### G. Segmentation Method

An approach to implement the Handwritten Character Recognition is provided by Ambadas B. Shinde and Yogesh H. Dandawate in 2014. This paper proposes a Word and Character Segmentation scheme for instrument printed Devanagari text. A complete word and character segmentation scheme for Devanagari printed text is presented in this. Sometimes, interline space and fused characters make line segmentation and character segmentation a hard task correspondingly. The authors have tested their method on...
documents in Marathi scripts. A novel technique of character segmentation for printed Devanagari text is presented in this. After removing the Shiorekha (header line) of Devanagari text, the bounding boxes are used to surround the segmented characters. Results obtained from this process are hopeful because of morphological operations. In this method the authors has proposed some basic morphological operations on the scanned document images and got much better results.

[14] 

H. K-Means Clustering

An approach to implement the Handwritten Character Recognition is provided by Akanksha Gaur and Sunita Yadav in 2015. In this paper recognition of Hindi characters are done by using a three step procedure. First step is pre-processing, in which binarization of the image and separations of characters are performed. Each Hindi word has a straight bar on the top of word. That bar is also eliminated in pre-processing phase. The next step is feature extraction in which region based k-means clustering is used and the feature vector is formed and used in classification stage as input. Third step is classification method, for which support vector machine in used. Support vector machine uses hyper-plane for categorization. This hyper-plane is used as a judgment surface which is with highest margin of separation of hyper-plane and closest data point. Support vector machine uses a different kernel functions which describes the mode of classification. The kernel function utilized in Support vector machine for classification is linear kernel function.

IV. OUTLINE OF THE HANDWRITTEN CHARACTER RECOGNITION IMPLEMENTATION TECHNIQUES

The following table contains all the techniques used to implement Handwritten Character Recognition that are explained previously and also contains their features, classifiers, findings and references number as well.

<table>
<thead>
<tr>
<th>Author &amp; Year</th>
<th>Features</th>
<th>Classifiers</th>
<th>Findings</th>
<th>Ref No</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Hanmandlu, O.V. Ramana Murthy, Vamsi Krishna Madaus</td>
<td>Box approach, Normalized Distance.</td>
<td>Fuzzy Sets.</td>
<td>Recognition Rate with Coarse Classification is 90.65%.</td>
<td>1</td>
</tr>
<tr>
<td>Raghuraj Singh, C. S. Yadav, Prabhat Verma Vibhash Yadav &amp; 2010</td>
<td>Mean Distance, Histogram of projection based on spatial position of pixel and pixel value.</td>
<td>ANN techniques</td>
<td>The input matrix of size 48X57 gives better results than other choices.</td>
<td>3</td>
</tr>
<tr>
<td>Pradeep Singh Thakur and Sandeep Patil &amp; 2012</td>
<td>means and covariance matrices of clusters</td>
<td>GHMM</td>
<td>The recognition percentage for different word sample is obtaining in between 30-60%.</td>
<td>6</td>
</tr>
<tr>
<td>Sajjad Pourmohammad, Reza Soosahabi, Anthony S. Maida &amp; 2013</td>
<td>MNIST dataset, Bilinear interpolation, PCA whitening, K-Means Clustering</td>
<td>Bayesian Classifier Using Mahalanobis Distance</td>
<td>Shows improvement as compared to other linear classifiers.</td>
<td>8</td>
</tr>
<tr>
<td>Swapnil A. Vaidya, Balaji R. Bombade &amp; 2013</td>
<td>Binary Image</td>
<td>GRNN classifier</td>
<td>The proposed recognition scheme provided 82.89% and 85.62% accuracies on Devnagari &amp; Kannada character databases Respectively.</td>
<td>7</td>
</tr>
<tr>
<td>Binny Thakral, Manoj Kumar &amp; 2014</td>
<td>Cluster detection</td>
<td></td>
<td>Overall 94.5% correctness for various input characters</td>
<td>10</td>
</tr>
<tr>
<td>Hann Meng and Daniel Morariu and 2014</td>
<td>Data Normalization</td>
<td>SOM, NN, MPN and Back propagation</td>
<td>Trained dataset results 65% correct predictions and Untrained dataset is 30% correct prediction with noise rate.</td>
<td>12</td>
</tr>
<tr>
<td>Mahesh Jangid, Sumit Srivastava &amp; 2014</td>
<td>ISIDCHAR applied ARAN &amp; V2DMDCHAR applied GLAC</td>
<td>SVM</td>
<td>The higher recognition accuracy is achieved 95.21% on V2DMDCHAR database.</td>
<td>13</td>
</tr>
<tr>
<td>Ambadas B. Shinde, Yogesh H. Dandawate &amp; 2014</td>
<td>Segment, Box approach</td>
<td></td>
<td>Shows 100% accuracy in formation of structural layout for document images as well as 100% accuracy for word segmentation.</td>
<td>14</td>
</tr>
<tr>
<td>Akanksha Gaur and Sunita Yadav and 2015</td>
<td>K-means clustering</td>
<td>Euclidean distance method, SVM</td>
<td>Result using Euclidean distance is 81.7% and SVM with linear kernel is 95.86%.</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 1: Accuracy of Various Recognition and Segmentation Techniques.

V. CONCLUSION

In this study, we have overviewed the different techniques in the character recognition domain attempting to bring out the present status of research in character recognition. Fuzzy sets classifier applied on box approach feature gives accuracy of 90.65%. The approach of GHMM classifier when used with features as means and covariance matrices of clusters then the resulting recognition percentage for different word sample is...
obtained in between 30-60%. GRNN classifier applied on Devnagari and Kannada character databases gives accuracies of 82.89% and 85.62% respectively by binary image. SVM gives the higher recognition accuracy of 95.21% on V2DMDCCHAR database applied with GLAC. K-Means Clustering using Euclidean distance and SVM with linear kernel gives accuracies of 81.7% and 95.86% respectively.

REFERENCES


[12] Hann Meng and Daniel Morariu, “Khmer Character Recognition using Artificial Neural Network” 2014 APSIPA.


