ISSN 2047-3338

A Review of Devnagari Character Recognition from Past to Future

Neeraj Pratap[1] and Dr. Shwetank Arya[2]
[1]BIT, MEERUT, India
[2]Gurukul Kangri University, India
[1]neerajpr23@gmail.com, [2]shwetank.arya@gmail.com

***Abstract*— In the last half century, the English character recognition was studied and the results were of such type that's it can produce technology driven applications. But the same approach cannot be used in case of Indian languages due to the nature of complication in terms of structure and computation. Now days there are different methodologies which are growing fastly in the area of Indian languages and character recognition. The offices, banks, schools and other organisations are working in the field of digital document processing. Devnagari is the national language of India and generally spoken by 600 million people in India. Devnagari should be given more special consideration for analysis and document retrieval due to its popularity. This paper is mainly concern for the people who are working in the Devnagari Optical character and it provides an overview about Devnagari character recognition system (DCR). In this paper the current status of DCR is presented and future research is also suggested.**

***Index Terms*— Image Classification, Segmentation, Devnagari Character Recognition, Feature Extraction and Thinning**

I. INTRODUCTION

NOW a days the character recognition has become a major area in the field of OCR (Optical Character Recognition). The contributions of Offline character recognition in the digital library evolution have found a great attention. In this paper we mainly focus on the handwritten Devnagri script. Handwriting recognition can be defined as the ability of a computer to translate human writing into text. The offline recognition process works on pictures which are generated by an optical scanner.

In this Paper the limitations of different methodologies is defined mainly based on two factors: one is the data collection process (On-line or off-line) and the second is type of text (hand written text or machine printed text). The Devnagari Character Recognition Problem can be defined based on five stages:

1. Pre-Processing
2. Segmentation
3. Feature Extraction
4. Recognition
5. Post processing

In this paper the different methodologies is defined on the basis of the five stages in the character recognition system. In this paper we have mainly focused on off-line character recognition. Although there are the different approaches for off-line and on-line character recognition but there also exists common solution for these two.

Due to the wide use of Pattern recognition and image processing technology, a lot of improvement have become in the handwritten recognition system. There are different types of pattern recognition techniques which can be helpful in the Devnagari optical character recognition.

In the Devnagari optical character recognition, the useful work is done by V.Bansal and R.M.K. Sinha [3]-[5]. The overview of statistical pattern recognition can be found in [1]-[2]. These references can be the best starting point to study the various types and applications of Devnagari Optical character recognition system. The general idea about the document analysis can also be found in [6].

In this paper the different characteristics of Devnagari language is defined in section 2. In section 3 the image pre-processing is defined, the segmentation stage is discussed in section 4.the different types of feature extraction is discussed in section 5 and the techniques for character classification is discussed in section 6. At last the post processing is defined in section 7 and finally the future research is defined in section 8.

II. FEATURES OF DEVNAGARI SCRIPT

The script Devanagari ('divine Nagari') is mainly based on phonologically, and it is written from left to right. If we study the history, like other native scripts of South Asia, it is derived from the Brahmi alphabet of the Ashokan inscriptions. Typologically, it is called as an alphasyllabary: that is, in it the consonant- vowel sequence is written as a unit, called as an aksara, in which the vowel symbol functions as an obligatory diacritic to the consonant.

In India Devnagari is considered as the most popular. In Devnagari there are 12 vowels and 33 consonants. These Vowels and Consonants are called as Basic Characters. The main features of vowels are that's it can be written as

independent letters, or by applying a methods of diacritical marks which are written above, below, before or after the consonant to which they belongs. The pattern of writing the vowels in this way is called as modifiers and the Characters which are made by this are called conjuncts. It can happen that's Sometimes that's two or more consonants can be combined and produce the new shapes. These types of new shape clusters are called as compound characters.

The other scripts also support these types of basic characters, compound characters and modifiers which are here in Devanagari. The national language of India "Hindi" is also in Devnagari script. Sanskrit, Nepali and Marathi language also use Devanagari for writing. In all over world the rank of Hindi is at number 3[2]. The character set which is used in Devanagari is shown in Table I to Table III.

Table I: Consonants

.

Table II: Vowels and their corresponding modifiers

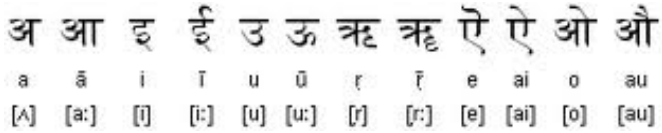
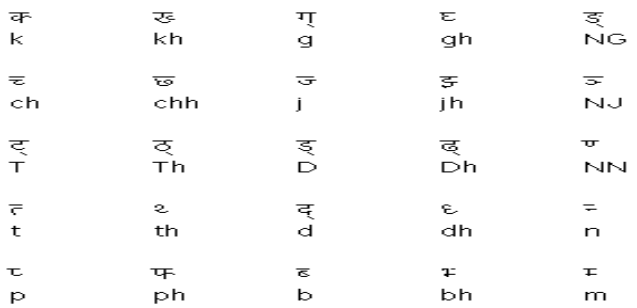


Table III: Half Form of the Consonants with Vertical bar



Shirorekha or headline is the horizontal line at the upper part of the characters. This characteristic is not found in English character so it can be taken as a distinguishable feature to extract English from these scripts. In the pattern of continuous handwriting from left to right, the Shirorekha of one character is attached with the Shirorekha of the previous or next character of the same word. So if we continue this manner, the multiple characters are appeared in a word as a single component having the common Shirorekha. In Devnagari there are a lot of characters which have the similar shape. There are also different vowels, numerals and compound characters due to these the Devnagrai optical character recognition have become a typical problem [7].

III. PRE-PROCESSING

The Paper document is generally scanned by the optical scanner and is converted in to the form of a picture. A picture is the combinations of picture elements which are also known as Pixels. The pixels contain basically two values ON and OFF. The ON value points that's the pixel is visible and the OFF value points that's the pixel is not visible. At this stage we have the data in the form of image and this image can be further analyzed so that's the important information can be retrieved. This analysis process can include the following points:

Noise Reduction
Normalization
Thresholding
Skew Detection
Thinning

A. Noise Reduction

When the document is scanned and is stored in picture format there is the chances that's the noise is introduced in the image. The noise can be introduced by optical scanning device or by the writing instrument. Due to the noise there can be the disconnected line segment , large gaps between the lines etc. so it is very essential to remove all of these errors so that's the information can be retrieved in the best way [8].

B. Normalization

The normalization is done so that's the character is arranged in a proper manner. Generally the individual character is fitted within the matrix. The matrix can be of 32 x 32 or 64 x 64 so that's the all characters can have the same size.

C. Thresholding

The thresholding is a process in which the gray scale image or any color image is converted into the binary image.

D. Skew Detection

For a document scanning process, there can be the skewness. The skewness should be removed because it reduces the accuracy of the document. The skew angle is calculated and with the help of skew angle, the skewed lines are made horizontal.

E. Thinning

To find the features of the objects, the boundary detection of the image is done. For boundary detection, various functions can be applied to the objects which are available in the MATLAB.

IV. SEGMENTATION

In Character Recognition techniques, the Segmentation is the most important process. Segmentation is done to make the separation between the individual characters of an image. The Devnagri words can be separated by removing Shirorekha. Each Separated character generates a sub-image. The Fig. 1 for Devnagrai segmentation can be shown as follows:

Fig. 1. Representation of Segmentation

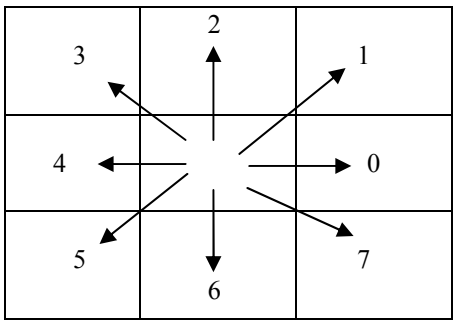
V. FEATURE EXTRACTION

Feature extraction is the process to retrieve the most important data from the raw data. The most important data means that's on the basis of that's the characters can be represented accurately. To store the different features of a character, the different classes are made. The feature of a character is stored to that's particular class to which it belongs [9]. So each class is different from another class. There are different methods which can be used for features extraction. These methods are classified mainly in the three groups.

- Statistical Features
- Geometrical and Topological Features
- Global Transformation and Series Expansion

A. Statistical Features

When the document image is represented by statistical distribution of point, the style variations of the characters is also considered. This is done so that's dimensions of the feature set can be reduced and also have the less complexity.



There are the different statistical features which can be used for character representation.

Zoning: The frame which contains the character is divided into different zones. The zones can be overlapping or non-overlapping zones. The different features or the point densities are analyzed in the different regions [10].

Projections: For the representation of Characters, the pixel gray value can be projected on to the lines in the different directions. By this representation the two dimensional image is converted to one dimensional signal which can be used for the representation of the character image.

Crossings and Distances: An important statistical feature is the number of crossing of a contour by a line segment in a specified direction. The character frame is divided into a set of regions in the different directions. So each region has the characters and the features of each region are extracted based on the characters.

B. Geometrical and Topological Features

A character can have the local and global properties. These properties can be represented by geometrical and topological structure with different style variations. These types of representation can provide help to extract the knowledge about the structure of the object or the knowledge about the basic components that make the object. The topological and geometrical representations can be categorized in the four groups.

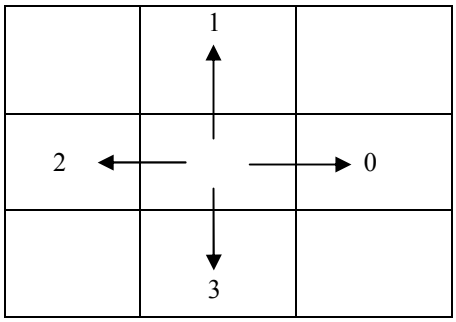
Measuring the Geometrical Properties: In this technique, the representation of the characters is done by calculating the geometrical quantities such as, the ratio calculation between height and width of the bounding box of a character, comparative lengths between two strokes, the relative vertical and horizontal distances between initial and last points, width of a stroke, the relative distance between the last point and the last y-min, word length curvature or change in the curvature, upper and lower masses of words [20].

Counting and Extracting the Topological Structures: In this type of structure, line ends, isolated dots, branch points, cross points, direction of a stroke from a special point, a bend between two points, straight strokes between two points, loops, maxima and minima, cups above and below a threshold and relationship between the strokes which constitutes a character are considered as the features [11].

Graphs and Trees: The Trees makes the hierarchical relation so this type of relationship can be used to represent the words or characters with a set of features. The words or characters can be partitioned into the different topological primitives like holes, crosspoints, strokes etc. the graphs can be used to represent the different primitives of a word or characters. The graph coordinates of the character shape can be used to represent the image.

Coding: For coding we can use Freeman chain coding. Freeman chain coding is used for detecting loops and curves in a character. The chain code can be used to represent the boundary by the connected sequence of straight line segments which is of specified length and direction. The numbering scheme can be used for the direction of each segment.

(a)



(b)

Fig. 2: (a) Eight Directional Version of Freeman Chain code, (b) Four directional Version of Freeman Chain code

C. Global Transformation and Series Expansion

It includes Fourier Transform, Gabor Transforms, Wavelets, moments and Karhuen-Loeve Expansion. The transformations or series expansion is the compact encoding which is provided by the linear combinations of the coefficients. Translation and rotation are the deformations under global transformation and series expansion.

Gabor Transform: The Gabor transform is the one type of short time Fourier transform. The use of Gabor transform is to find the sinusoidal frequency. The Gabor transform is also used to find the phase content of local sections of a signal as it changes over time. The function which is to be transformed first of all is multiplied by a Gaussian function, and the result is known as a window function. The window function is then transformed with a Fourier transform which derives the time-frequency analysis. The window function means that the signal near the time being analyzed will have higher weight.

Wavelets: Wavelet transformation is used to represent the signal at different levels of resolution. Corresponding to various levels of resolution the wavelets coefficients are used to represent the segments of document image correspond to letters or words. These coefficients work as an input to a classifier for recognition purpose.

Fourier Transforms: The relationship between a signal in the time domain and its representation in the frequency domain is defined by the Fourier transform. The most unique feature of Fourier transform is to recognize the position-shifted characters, when it observes the magnitude spectrum and ignores the phase [8].In Devnagri, Fourier Transforms has been applied in many ways [12].

Karhunen-Loeve Expansion: Karhunen-Loeve expansion is the eigen-vector analysis. Karhunen-Loeve Expansion tries to decrease the feature set by creating the new feature set. The new feature set is the linear combinations of the original ones. There are different pattern recognition problems like face recognition where Karhunen-Loeve expansion is used. In character Recognition problem Karhunen-Loeve expansion is not used widely because complex algorithms are required to implement it.

VI. CLASSIFICATION OF CHARACTER

Optical Character Recognition widely use the pattern Recognition techniques. There are also the different techniques which are investigated by the researchers. A survey report on feature extraction and classification methods for Devnagari character recognition can be found in [8].

There are the different classification techniques for Optical character Recognition.

- Neural Networks.
- Statistical Techniques.
- Template Matching.
- Support Vector Machine (SVM) algorithms.
- Combination classifier.

The above techniques are not purely independent. The one technique can be considered as the subset of another.

A. Neural Network

To develop an accurate OCR system is a complicated task and requires a lot of effort. Such types of systems are usually complicated and can hide a lot of logic behind the code. To achieve the good performance and improved quality of recognition, the artificial neural network in OCR applications performs the important function. Another benefit of using neural network in OCR is extensibility of the system – ability to recognize more character sets than initially defined. There are the different types of neural networks which can be used for the character recognition.

The artificial Neural Network is a computing architecture which consists of 'neural' processors connected in a parallel sequence. The Artificial Neural Networks can perform the computation at a higher rate as compared to the classical techniques because it has the parallel nature. The output from one node goes to another node as the input and the final decision depends on the complex interaction of all nodes. The neural network can be categorized in to two types, feed forward and feedback network. The feedback network is also known as the recurrent network.

In OCR system, the most common neural network is multilayer perceptron (MLP) network which is of type feed forward network. The interesting Feature of MLP is that it provides the confidence in the character classification. First of all MLP is proposed by U. Bhattacharya et al [13]. M.Egmont-Petersen has shown the comparison of various NN classifiers like Feed forward, Neuro-fuzzy system etc. for English language, K.YRajput et al[14] have also used classifier like back propagation type which is based on Genetic algorithm and also classification along with fusion of NN and Fuzzy logic.

B. Statistical Techniques

Statistical decision theory is based on the statistical decision functions and a set of optimality criteria, which maximizes the probability of the observed pattern given the model of a certain class. Statistical techniques are based on following assumptions:

a. The Distribution of the feature set is based on Gaussian or in the worst case uniform.

b. There are a wide statistics available for each class.

c. With the collection of images, a set of features can be extracted which represents each distinct class of patterns. The value taken from n-features of each word unit can be consider representing an n-dimensional vector space and the vector, whose coordinates correspond to the value taken and represents the original word unit. The main statistical methods which are applied in the OCR field are Clustering Analysis [15], Hidden Markov Modeling (HMM), Nearest Neighbor (NN) [13], Likelihood or Bayes classifier, Fuzzy Set Reasoning, and Quadratic classifier.

C. Template Matching

Template matching is one of the Optical Character Recognition techniques. Template matching is the process of finding the location of a sub image called a template inside an image. Once a number of corresponding templates is found their centers are used as corresponding points to determine the registration parameters. Template matching determines the similarities between a given template and windows of the same size in an image and identifying the window that produces the highest similarity measure. The recognition rate of Template Matching is mainly depending on noise and image deformation. For improved classification Deformable Templates and Elastic Matching are used.

D. Support Vector Machine Classifier

For Data Classification, SVMs (Support Vector Machines) are a useful technique. A classification process generally involves separating the data into two sets, training and testing sets. Each instance in the training set contains one target value (i.e. the class labels) and several Attributes (i.e. the features or observed variables).The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. Many researchers used SVM successfully viz. Umapada Pal et al. [16], Sandhya Arora et al. [13].

E. Combination Classifier

To improve recognition results, decisions of different classifiers can be combined. The combination can be done in different ways which depends on the types of information produced by the individual classifiers. Different approaches to combine multiple classifiers can be classified into three main categories according to their architecture: 1) cascading (or serial combination) 2) parallel, and 3) hierarchical (tree-like).

Selection of Individual Classifiers: If the individual classifiers are independent in a great extent, the classifier combination is especially useful. If this is not done by the different training sets, various recompiling techniques such as bootstrapping and rotation can be used to create differences for improving the classification rate.

Combiner: The Classifiers needs to be combined together when they are selected. The combination is done by a module, called as the combiner. The combiners can be different from each other in their adaptivity, trainability, and requirement on the output of individual classifiers. Combiners, like to do voting or to perform averaging (or sum), are static. For these static operations no training are required. On the other side the trainable combiners may provide the best result than to the static combiners at the cost of additional training as well as the requirement of additional training data.

Some combination classifiers used in Indian scripts are SVM and ANN [13], ANN and HMM [17], MLP and SVM [17], MLP and minimum edit [17], K-Means and SVM [18], NN, fuzzy logic and genetic algorithm. The five classifiers (three NN based and two HMM) are used by Pavan Kumar [19] to obtain the accuracy at the best level.

VII. POST PROCESSING

The objective of preprocessing is to clean the document in a certain sense but the important information can be removed because there is the lack of context information at this stage. Generally the complete OCR problem is to determine the context of the document image. So the combination of context and shape information in all the phases of OCR systems is necessary for meaningful improvements in recognition rates. The accuracy of Character Recognition stages can be improved if the semantic information were available upto a great extent. In the Post processing stage, this is done with the feedback to the previous stages of OCR. The best way of incorporating the context information is the use of dictionary which can be used to correct the small mistakes of OCR system. The main idea for spell checking the OCR output and also provides some techniques for the output of the recognizer which are not in dictionary. For Devnagari Language, the Research is underway to check the spelling. When the unknown character is recognized, it can be saved.

VIII. FUTURE RESEARCH

India is a country which have multilingual and multiple-script so Research and development in Indic language processing is a necessity. A Program on Technology Development for Indian Languages (TDIL: http://www.tdil.mit.gov.in) have been started by Ministry of Information Technology of Government of India, where language aspects are studied and developed. CDAC (Centre for Development of Advance Computing) is also actively involved in development of Indian languages. Various language translators are developed by CDAC in collaboration with IIT Kanpur for word processing. IIT Kanpur, IIIT Hydrabad, ISI Kolkata and many others Research institutes are working in Devnagari character Recognition [8].

OCR has been investigated for the different Indian scripts like Tamil, Bengali, Telugu, Gurumukhi etc. Most of the research is based on to identify the isolated character rather

than the word, phrases, or the entire document. The Indian scripts are a composition of the constituent symbols in two dimensions. In conventional Research, first the segmentation process is applied to word so that's the word is segmented into its composite characters. Later on the composite character is decomposed into symbols or strokes (like Matra) which are finally recognized. A lot of research is still needed for word, sentence and document recognition, its semantics and lexicon. There is still a need to do the research in the area Devnagari character recognition.

IX. CONCLUSION

In the last two decades the different methods have been proposed by the Researchers for treating the problem of Devnagari character. But a lot of research is also required to handle the Challenges in Devnagari Character Recognition. It is hoped that this detailed discussion will be beneficial insight into various concepts involved, and boost further advances in the area. The accurate recognition is directly depending on the nature of the material to be read and by its quality. Current research is not directly concern to the characters, but also words and phrases, and even the complete documents. For the character recognition, HMM, neural networks and their combinations are used as the powerful tools. For the high reliability in character recognition, segmentation and classification have to be treated in an integrated manner to obtain more accuracy in complex cases. This paper has focused on an appreciation of principles and methods. The compare of effectiveness of various algorithms has not been attempted in the present work. Unfortunately there is little experimental as well as standard handwritten character database available publicly for benchmarking the accuracy of various advanced techniques proposed in Devnagari character recognition.

To understand the approaches described in more details, the list of references is enlisted. We thanks to researchers whose important contributions may have been overlooked.

REFERENCES

[1] R.G. Casey, D. R. Furgson, *"Intelligent Forms Processing",* IBM System Journal, Vol. 29, No. 3, 1990.

[2] U. Pal, B. B. Chaudhuri, *"Indian Script Character recognition: A survey"*, Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

[3] R.M.K. Sinha and Veena Bansal, "*On Automating Trainer For Construction of Prototypes for Devnagari Text Recognition"*, Technical Report TRCS-95-232, I.I.T. Kanpur, India.

[4] R.M.K.Sinha., *"Rule Based Contextual Post-processing for Devnagari Text Recognition*", Pattern Recognition, Vol. 20, No. 5, pp. 475-485, 1987.

[5] R.M.K.Sinha, "*On Partitioning a Dictionary for Visual Text Recognition*", Pattern Recognition, Vol 23, No. 5, pp 497-500, 1989.

[6] Rangachar Kasturi, Lawrence O'Gorman, Venu Govindaraju, "*Document Image Analysis: A Primer*", Sadhana, Vol. 27, Part 1, pp. 3–22, February 2002.

[7] Ram Sarkar et al, "*A Script Independent Technique for Extraction of Characters from Handwritten Word Images*", International Journal of Computer Applications Vol. 1, No. 23, 2010.

[8] Nafiz Arica, Fatos T. Yarman-Vural, "*An Overview of Character Recognition Focused On Off-line Handwriting*", C99-06-C-203, IEEE, 2000.

[9] I. S. Oh, J. S. Lee, C. Y. Suen, "*Analysis of class separation and Combination of Class-Dependent* Features for Handwriting *Recognition*", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.21, no.10, pp.1089-1094, 1999.

[10] S.V. Rajashekararadhya, P. Vanaja Ranjan, "*A Novel Zone Based Feature Extraction Algorithm for Handwritten Numeral Recognition of Four Indian Scripts*", Digital Technology, Journal, Vol. 2, pp. 41-51, 2009.

[11] D. Trier, A. K. Jain, T. Text, "*Feature Extraction Method for Character Recognition - A Survey*", Pattern recognition, vol.29, no.4, pp.641-662, 1996.

[12] G. G. Rajput, S. M. Mali," *Fourier Descriptor based Isolated Marathi Handwritten Numeral Recognition*", Int'l Journal of Computer Applications, Volume 3 – No.4, June 2010.

[13] Sandhya Arora et al., "*Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition*", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, May 2010.

[14] K. Y. Rajput and Sangeeta Mishra, "*Recognition and Editing of Devnagari Handwriting Using Neural Network*", SPIT-IEEE Colloquium and Intl. Conference, Mumbai, India.

[15] Judith Hochberg, Lila Kerns, Patrick Kelly, and Timothy Thomas, "*Automatic Script Identification from Images using Cluster-based Templates*", IEEE transactions on PAMI, Vol.19, issue-2 pp 176 – 181, 1997.

[16] Umapada Pal, Sukalpa Chanda Tetsushi, Wakabayashi, Fumitaka Kimura, "*Accuracy Improvement of Devnagari Character Recognition Combining SVM and MQDF*".

[17] U. Bhattacharya, S. K. Parui, B. Shaw, K. Bhattacharya, "*Neural Combination of ANN and HMM for Handwritten Devnagari Numeral Recognition*".

[18] Satish Kumar, "*Evaluation of Orthogonal Directional Gradients on Hand-Printed Datasets*", Intl. Journal of Information Technology and Knowledge Management, Volume 2, No. 1, pp. 203-207. Jan - Jun 2009.

[19] M N S K Pavan Kumar, S S Ravikiran, Abhishek Nayani, C V Jawahar, P J Narayanan , "*Tools for Developing OCRs for Indian Scripts*".

[20] Vikas J Dongre, Vijay H Mankar, *A Review of Research on Devnagari Character Recognition*, International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010.