A Review on Various Techniques of Devanagari Script Recognition

Pooja Yadav[#1], Sonia Sharma[*2]

[1]M.Tech., CSE Department, JMIT, Radaur, Kurukshetra University, India
[2]Assistant Professor, CSE Department, JMIT, Radaur, Kurukshetra University, India

Abstract *— Image processing has wide area for processing various functionality on image. Image with any pattern comes under the categories of pattern recognition where recognizing of the pattern can be any character, symbol, numeral or it can be any image also. Character Recognition (CR) has broad area of research in English as well as Hindi character. English language has been largely progressed to a level, sufficient to produce technology based applications. But Hindi languages are complicated structure and their computations so this language has not progressed well. Devanagari character recognition provides less correctness and efficiency. To recognize Hindi Devanagari script, various development done which is discuss in detail. Developers used to recognize the pattern with their structure, template, and graph. Some developers use classifiers to segmenting the characters.*

Keywords — *Devanagari Script, Offline Character recognition, Feature Extraction, Segmentation, Neural Network*

I. INTRODUCTION

A. *Pattern Recognition*

It is defined as the field concerned with machine recognition of meaningful regularities in noisy and complex environments. Pattern recognition techniques associated symbolic value with the image of the pattern. Pattern recognition applicable in character recognition, online signature verification, and face recognition and so on

Four significant approaches to PR have evolved. Which are Statistical based, Syntactic based, neural network and Knowledge-based [1].

B. *Character Recognition*

Character recognition is part of pattern recognition field in which images of characters from a text image are recognized and recognition result as character codes are returned [10]. It is the process of recognizing typed, printed or handwritten characters and converting into machine readable code. Process of converting printed or handwritten scanned documents into their corresponding ASCII characters that system can recognize converting image of documents into digital textual equivalent. Character recognition can be used for automatic number plate recognition, converting handwriting in real time to control a computer, as a reading aid for blind etc.

1) Offline System:

It is based on the type of the text which is printed or hand written. In this type of character recognition as handwritten, type written or printed text is well transformed into digital format. There does not exist benefits of recognizing direction of the movements while writing the character. The typewritten or handwritten character is normally scanned in form of paper document and store it in codes of a binary/gray scale image to the recognition algorithm.

2) Online System:

It is the two dimensional coordinates represented as function of time and order of strokes. The on-line methods superior to off-line in recognizing handwritten characters as temporal information available.

Online character recognition is much easier and achieved better results than offline character recognition because more information may be captured in online like direction, speed and order of strokes of the handwriting [9].

II. DEVANAGARI SCRIPT

A. *Devanagari script*

It is the basic script of various languages which are speaking in India, such as Hindi and Sanskrit. This script is composition of symbols in two dimensions. In script, horizontal writing style, from left to right and characters do not have any uppercase/lowercase. Devanagari is a phonetic and syllabic script. Devanagari is phonetic, as words are written exactly as they are pronounced and syllabic means that text is written using consonants and vowels that together form syllables. Devnagari script has 13 vowels and 36 consonants and 11 modifiers [6].

Devanagari has some features as:

B. *Vowels*

The vowels use in two ways in both English and Hindi: to produce their own sounds and to modify sound of consonant; too this, an appropriate modifier

from the symbols is mixed with consonants in right manner. Vowels occur either in isolation or in combination with consonants.

C. Consonants

Consonants are exists in a pure form called a half character for almost every consonant. It may be possible that pure form have combined with other consonants. In normally, the half forms of consonants, with the right part removed, are placed with left part of original consonants. Some special characters may appear in the lower half of the new composite forms [4].

Vowels [स्वर]	अ	आ	इ	ई	उ	Modifiers	ा	ि	ी	ु
	[1]	[2]	[3]	[4]	[5]		[1]	[2]	[3]	[4]
	ऊ	ए	ऐ	ओ	औ		ू	े	ै	ो
	[6]	[7]	[8]	[9]	[10]		[5]	[6]	[7]	[8]
	अं	अः	ऋ				ौ	ं	ृ	
	[11]	[12]	[13]				[9]	[10]	[11]	
Consonants [व्यनजन]	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
	ट	ठ	ड	ढ	ण	त	थ	द	ध	न
	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
	प	फ	ब	भ	म	य	र	ल	व	श
	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]	[29]	[30]
	ष	स	ह	ळ	क्ष	ज्ञ				
	[31]	[32]	[33]	[34]	[35]	[36]				

Fig. 1 Consonants, vowels modifiers [7]

D. Modifiers

A pure and full consonant with a modifier becomes a modified character, which is then concatenated with other basic/modified characters to form a word. These symbols are placed next to the consonants (core modifier), above consonants (top modifier) and below consonants (lower modifier) the consonants. Some modifiers contain core and top modifier as core modifier placed before or next to the consonant and top modifier placed above the core modifier.

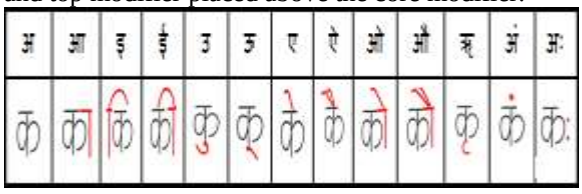

Fig. 2 Consonants with modifiers [7]

E. Consonant Forms

There are various forms of consonant which are given below:

1) Conjunct Characters:

The next character is always touches the consonant which is in the pure form the occurring in the script and then this characters called as conjuncts, touching characters or fused characters.

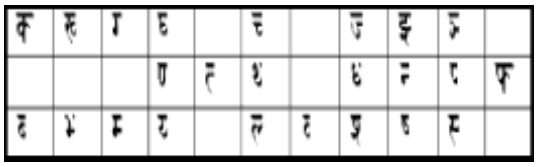

Fig. 3 Pure form of Consonant [5]

A character written with its half form and then attached to another basic character, which is complete. This combination is called as conjunct character.

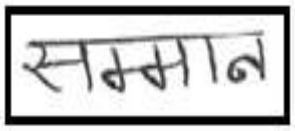

Fig. 4 Conjunct Character [6]

2) Overlapping Characters:

Two or more characters in a word overlap with one another due to rushed or hasty handwriting of a few writer.

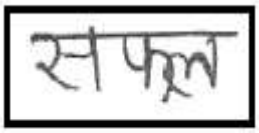

Fig. 5 Overlapping Character [6]

3) Other Forms:

There also exist some characters having half forms of consonants are the left part of original consonants with the right part removed. A consonant in pure form always touches the next character occurring in the script. Some special combinations where a new character or the half forms of consonants may appear in the lower half of the new composite forms [6].

Fig. 6 combinations of consonant [5]

F. Header Line

The header line is a horizontal line which is drawn at the top of each character and extends throughout the word in Devanagari. The presence of a horizontal line on the top of all characters. This line is known as header line or "Shirorekha". For various specific characters, the Shirorekha is written only partially on top of characters.

Four lines of words are:

1) Head line: This line same as header line, separates the top and core strips

2) Virtual Baseline: This line exist where characters completed but excluding below modifiers. It separates the core and lower strips.

3) Lower line: Drawn below modifiers line

4) Upper line: Line above Headline after modifiers

Three zones of words are:

1) Upper zone: Between Upper line and Headline,

2) Middle zone: Between Headline and Baseline

3) Lower zone: Between Baseline and Lower line [10].

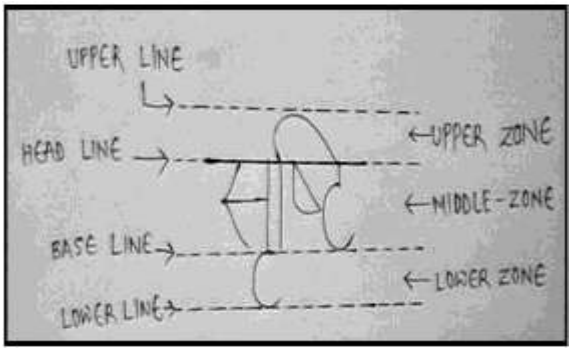

Fig. 7 Lines and zones of a word [6]

III. Recognition Of DevAnagari Script

Document which want to recognize, is scanned by optical scanner and is converted in to the form of a picture. Picture exist various combinations of picture elements called as Pixels. These pixels contain basically two values ON and OFF. The ON value pixels are the pixels which are visible and the OFF value points that's the pixel is not visible.

No matter in which class recognition document will belongs, generally, there are four major stages in recognition process:

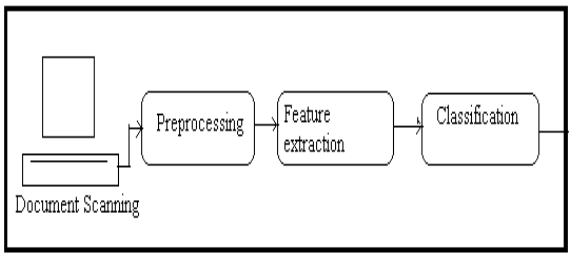

Fig. 8 Steps for Recognition [4]

A. Pre - Processing

This is the first stage where we have data in the form of image and in this stage, the image can be used for more analysis by which the required information can be retrieved for further stages. Pre-processing includes various processes which required for shaping of the input image into a required form.The pre-processing is a series of operation performed on the scanner input image. Through this step, whole enhancement of the input image features done at this level at which make suitable for segmentation.

1) Binarization Process:

It converts a gray scale image into binary image using thresholding technique. This thresholding refers to the conversion of a gray-scale image into a binary image. For conversion of gray level image into binary form, there exist mainly two approaches.

i. Global Threshold:

It picks only one threshold value from whole entire image, which based on estimation of the background level from the intensity histogram of the image.

ii. Local Threshold:

It uses different values for each pixel according to the local area information

2) Noise Elimination:

Various noise can be exists in images which make one of the major issue in pattern recognition process. Noise factor plays vital role for quality of images, as quality reduce with increasing of noise. Noise can be occur at different stages of

recognition like image capturing, transmission and compression. It is also called smoothing. Smoothing is basically used to reduce the fine textured noise and for improve the quality of the image.

3) Size Normalization:

Normalization is applied on scanned document to obtain all characters with uniform size. It provides a tremendous reduction in data size. As in document character patterns written with different sizes. The array with a fixed size use as a input to the neural network is an. For setting of the image size which suitable to network size, size normalization is required. Normalization mainly use to reduce the size of the whole image without requiring any of the structure information of the image.

4) Thinning:

Thinning is used to remove selected foreground pixels from binary images. Thinning always extracts the features related shape information of the characters. This thinning process also reduces the memory space required for storing whole information of whole input characters and it also reduces the processing time too. This process can remove irregularities in letters and makes algorithm simpler because other stages have to operate on the single character stroke, which is only one pixel value wide area [4].

B. Segmentation

The Segmentation is the most important and enhancing process. Mainly, segmentation is done to make the separation between single characters of previous stage image. Related Devanagri words can be make separate by removing the header line. Each Separated character generates a sub-image. Sub image is segmented into isolated characters by assigning a number to each character through the labeling process. This labeling process provides the information about the number of characters in the image and make each character resized into m x n pixels which suitable for training neural network. Character segmentation, is an operation that use to decompose whole image of sequence of characters into their sub images of the individual symbols. This is one of the main decision process in a recognition system for character recognition. Its decision about right or wrong of pattern isolated from the image can be treated as error report.

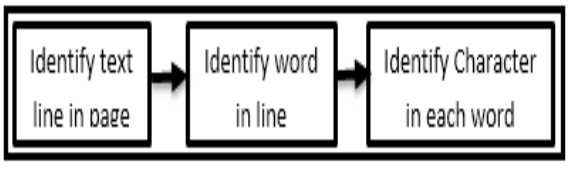

Fig. 9 Steps for Segmentation

C. Feature Extraction

Both feature extraction and selection comes intentionally to extracting the final representative features from the previous raw data, through which minimizes the features related to within class pattern variability as well as enhancing features between class pattern variability. For this purpose, various set of features are extracted from each class which further helps in distinguish it from other classes, while remaining invariant to characteristic differences within the feature class.

There exist many features in image which can be extracted for the recognition of Devanagari characters for that we consider features as follows:

1) Histogram plot:

The histogram is mainly use for the graphical representation of distribution of data. In graphs, x-axis shows the range of values in Y and other y-axis shows the number of elements that fall within the groups; y-axis ranges from 0 to the greatest number of elements .The x-range of the leftmost and rightmost bins can be extends for represent the entire data range in the case when the user-specified range does not cover the whole data range; as give results in "boxes" from both edges of the distribution.

2) Gray Level Co-occurrence Matrix (GLCM):

This is the statistical method for examining the texture which consider for the spatial relationship between the pixels called as the gray-level co-occurrence matrix (GLCM), this is also called as gray-level spatial dependence matrix. The GLCM functions is mainly use for characterize the texture of whole image by calculating how often the pairs of pixel with specific values is achieved and in a specified spatial relationship occur in an image, then it creating a GLCM, and then it extracting statistical measures from this matrix.

3) Color Domino:

Surf and surfc functions are mainly required for the color parametric Surfaces being specified over the X, Y and Z axis.

The selection of appropriate feature extraction method play the single most important factor in achieving high performance and high accuracy [10].

D. Classification

Last decision making stage of an OCR system is the classification stage and uses the features extracted in previous stage to identify text segment according to preset rules.

The feature obtained from previous phase is assigned a class label and in this stage whole feature recognized by using supervised and unsupervised

techniques. The whole data set can be divided into two sets as training set and test set for each character.

Classification is performed based on the extracted features which extensively use the logics of pattern recognition by assigning some unknown samples of the input to a predefined class of network various techniques for recognition are reviewed by the researchers. OCR classification techniques mainly categories as follows:

1) Template Matching:

It involve the computation class of conditional probability densities. Techniques are Direct matching, deformable and elastic matching, relaxation matching

2) Statistical Techniques:

This technique uses rule-based system with artificial intelligence (AI). In this rule-based system, each rule is in format of a clause which reflects the evidence about the presence of a particular class. Techniques are parametric, Non-parametric, HMM, fuzzy set reasoning.

3) Structural Techniques:

In this each pattern is composed of sub-pattern with the generation rules of a grammar characteristic. Class identifications done by parsing operations using automata corresponding to the various grammars. Techniques are graphical methods and grammatical methods.

4) Neural Networks:

Network has large number of neurons connected by weighted links. Character recognition. Techniques are Multilayer perceptron, radial basis function and support vector machine.

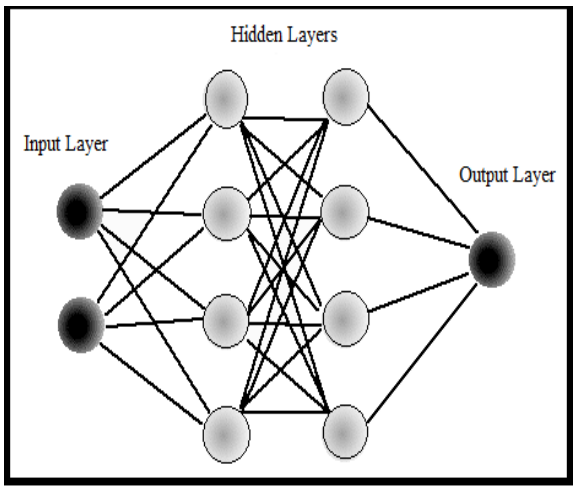

Fig. 10 Architecture of Neural Network

5) Hybrid or Combination Classifier:

Combination scheme constructed by the set of individual classifiers and combiner which combines the results of each individual classifiers for making whole single and final decision. There are various schemes for combining multiple classifiers can be grouped into three main categories which based on their architecture parallel, cascading, and hierarchical.

6) Support Vector Machines (SVM):

The SVM is mainly use to map the whole input data into a higher dimensional feature space which is mostly nonlinearly and it related to the input space which use to determine the separated hyper plane which have the maximum margin between the two classes in the feature space. We can built this maximal margin of hyper plane in feature space by using the kernel function in gene space operator. We can be determine the optimal separating hyper plane without any computations of higher and superior dimensional feature space with the help of using kernel functions in the input space of plane [2].

IV. Related Review Work

As through Review various literature it clear that first research on handwritten Devanagari characters was published in 1977 from this there is not much research work is done after that. At present, we found that many researchers had perform some development towards the off-line handwritten Devanagari characters and in a meanwhile, few research reports are published recently & tried to solve the problem associated with them [6].

A. Feature Extraction

There are lots of methods for feature extraction in character recognition that have been reviewed. Developers used different feature extraction methods in their research as Template matching, Deformable templates, Graph description, Projection Histograms, Contour profiles, Fourier descriptors, Gradients feature.

Anshul Gupta, Manisha Srivastava and Chitralekha Mahanta [2] represented two approaches as holistic and segmentation. Holistic approach used in recognition of limited size vocabulary where global features extracted from the whole word image. When size of the vocabulary increases, then complexity of holistic based algorithms increases and recognition rate decreases. They used segmentation based strategies, bottom-up approaches, where stroke or the character level considered and producing a meaningful word.

Dr. Latesh Malik [5] used a graph based holistic approach which based on sub graphs homo-morphism to recognition for handwritten Devnagari OCR and it represent prototypes of words. There exist Feature graph which extracted from input word. It is extended to net which is created by new nodes. Each sub graph which recognized will act as

node in a directed net this node compiles different features in the feature graph.

Shruti Agarwal and Dr. Naveen Hemarjani [6] focused on the recognition of offline handwritten Hindi characters by used of template matching algorithm in devnagari script. Characters are mainly work as document images in OCR. Handwriting recognition is the property of the system to receive and interpret handwritten input of photographs, touch-screens, paper documents and other devices.

Gayathri P and Sonal Ayyappan [9] gave method for recognition of Malayalam handwritten vowels using Hidden Markov Model (HMM). OCR for Recognition of handwritten Malayalam vowels is discussed. Using Hidden Markov Model Toolkit, Training and recognition are performed [9].

Ashutosh Aggarwal, Rajneesh Rani and Renu Dhir [3] described novel methods of recognition of single isolated Devanagari character images through feature extraction approach. He used this approach which is flexible in that the same previous algorithms, without modification, for feature extraction in a variety of OCR problems like handwritten, machine-print, gray scale, and binary and low-resolution character recognition. The gradient representation shows basis for extraction of features. Algorithms require a few simple arithmetic operations per image pixel which suitable for real-time applications.

B. Genetic Algorithm

There exist various methods of feature extraction for character recognition using genetic algorithm have been reviewed. Genetic programming (GP) is the fast developing approach to automatic programming in recognition of characters. In genetic programming, there are various solutions to a problem which are represented as a programs in computer. Developer use Darwinian principles for natural selection and recombination.

Prof. Mukund R. Joshi and Miss. Vrushali V. Sabale [10] presented a guide and update for the readers, working in the Devanagari Character Recognition area, whole knows that a single classification algorithm never gives better performance rate, so Author use not only neural network but also genetic algorithm.

Ved Prakash Agnihotri [4] described Diagonal based feature extraction is used for extracting features of the handwritten Devanagari script. Features of each character image is converted into chromosome bit string of length. Diagonal based feature extraction method extract 54 features from each character. Character recognize image in which extracted features are converted in Chromosome bit string of size 378. In recognition step using fitness Function where it will find differences in unknown Character and Chromosome which store in database.

C. Neural Network

Several methods of artificial neurons and their processes information. Developers used connection Approach to computation Modern neural networks by implemented non- linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs and then form this relation find patterns in data..

Yash Pal Singh, Abhilash Khare and Amit gupta [1] analyzed neural network method in pattern recognition. He focused on solutions which applied Hopfield Auto associative memory model for pattern recognition. This network is an associative memory. The primary function of this memory is to retrieve the pattern which stored in memory, when there is an incomplete or noisy version of that pattern is discussed.

Divakar Yadav , Sonia Sánchez-Cuadrado and Jorge Morato [8] proposed an OCR for printed Hindi text in Devanagari script, by used techniques of Artificial Neural Network (ANN), and then improved recognition efficiency. He performed conversion of gray scaled images to binary images and a back-propagation neural network having two hidden layers is used.The classifier is trained and tested for printed Hindi texts.

S Sayyad , Manoj Jadhav, Smita Miraje, Pradip Bele and Avinash Pandhare, [7] used Devnagari characters using multilayer perceptron with hidden layer. Various patterns of characters created in matrix (n*n) in binary form and stored in the file. Author used the back propagation neural network for efficient recognition with rectified neuron values of feed forward method in the neural network.

There has been a dramatic increase of research in the field of Devanagari OCR since 1990. Various recognition techniques which used by recognition are depend on the nature of the data to be recognized.

Few research reports are published recently & tried to solve the problem associated with them. We found that many researchers had done work towards the off-line handwritten Devanagari characters

Review of all developers research being done and main factors of their research is being tabulated in the Table 1.In this table, there exist developer's paper published year, technique which author used, merits and demerits of their techniques, recognition accuracy with their technique and whole result of their research.

TABLE I

REVIEW OF CHARACTER RECOGNITION TECHNIQUES

Year	Technique	Accuracy	Merits	Demerits	Result
2010	Hopfield Auto associative memory Model [1]	98%	Recognizes well when input vector equal to the stored patterns and more random error. It also eliminated the noise	Limited capacity. It difficulty to the recovery of stored patterns if it closed to its hamming distance.	It work as recurrent neural processing within parallel architecture.
2011	Holistic and segmentation based artificial neural network [2]	98.74%	Efficient method increase performance. Heuristic algorithm are deduced empirically and improve efficiency of the system.	No guarantee about optimum results for Different styles of writing.	Optimal results of character recognition technique. .
2012	Graph based holistic approach [5]	84%	Net can easily manage with additional marks about occurrence of symbols	Research analysis on more calculation of matching quality Is required	It recognize sub graphs as meaningful units
2012	Diagonal based feature using genetic algorithm [4]	85.78%	Enhance accuracy by using genetic attributes like Chromosome bit string of 54 feature	Need more complex method to optimize the solution	Best result conducted on training set of 904 characters and testing set of 204 characters.
2012	Templates matching [6]	92.66%	Good accuracy with noise free scanned images.	Documents must be written by same person	System tested and accuracy measured for each templates.
2013	Back propagate, feed forward method neural network [7]	86%	Different orientation, thickness use to generalize the missing data	Character is recognized within predefined range only	Efficient recognition with good accuracy
2013	Histogram base projection mean Distance, zero crossing. [8]	90%	Powerful techniques for extract features of distorted characters/symbols	It does not dealt with various punctuation marks and more numerals	Improve the rate of character recognition accuracy. .
2013	Gradient Vector Feature [3]	94%	Logical simplicity, easy use and high recognition rate, Gradient Features	Any change in γ (gamma) decreases the recognition rate	It becomes stable at higher values of Cost parameter recognition rate
2014	Hidden Markov Model (HMM) [9]	81.38%	Remove the redundant features from the feature space and thus improve accuracy.	Difficult to perform Markov Model Tool kit operation	Markov increase accuracy and Training recognised Malayalam characters
2015	Neural network and genetic algorithm [10]	85%	Increase the efficiency by segmentation with Genetic algorithm and neural network	Recognition was not up to expectations because shape of Characters or quality of image.	Overall successful recognition achieved is better than the previous result.

V. CONCLUSION

Various recognition techniques which used by recognition are mainly depend on the nature of the data to be recognized. There has been a dramatic increase of research in the field of Devanagari OCR since 1990. Various strategies are being used by the developers in reviewed. As holistic approach give good accuracy but it done recognition of limited size vocabulary. Segmentation based strategies, have no any limitation of size but there is no any guarantee about optimum results for different styles of writing. By using Hidden Markov Model (HMM), it remove the redundant features from the feature space and thus improve accuracy but it difficult to perform Markov Model Tool kit operation. In graph based holistic approach net can easily manage with additional marks about occurrence of symbols but Research analysis on calculation of matching quality is required. In back propagate, feed forward method neural network by using different orientation, thickness use to generalize the missing data but character is recognized within predefined range only.

REFERENCES

[1] Yash Pal Singh, Abhilash Khare and Amit gupta," *Analysis of Hopfield Autoassociative Memory in the Character Recognition*", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 500-503

[2] Anshul Gupta, Manisha Srivastava and Chitralekha Mahanta," *Offline Handwritten Character Recognition Using Neural Network*", International Conference on Computer Applications and Industrial Electronics (ICCAIE 2011), 2011

[3] Ashutosh Aggarwal, Rajneesh Rani and RenuDhir,"*Handwritten Devanagari Character Recognition Using Gradient Features*", International Journal of Advanced Research in Computer Science and Software Engineering Research Paper,Volume 2,Issue 5, May 2012, ISSN: 2277 128X

[4] Ved Prakash Agnihotri,"*Offline Handwritten Devanagari Script Recognition*", 378-1-3299-3020-6190/$31.00 ©2012 IEEE, 2012

[5] Dr. Latesh Malik,"*A Graph Based Approach for Handwritten Devnagari Word Recognition*", Fifth International Conference on Emerging Trends in Engineering and Technology, 2012

[6] Shruti Agarwal and Dr. Naveen Hemarjani,"*Offline Handwritten Character Recognition with Devnagari Script*", IOSR Journal of Computer Engineering (IOSR-JCE) ,Volume 12, Issue 2 (May. - Jun. 2013), PP 82-86, e-ISSN: 2278-0661, p- ISSN: 2278-8727

[7] S S Sayyad, Abhay Jadhav, Manoj Jadhav, Smita Miraje, Pradip Bele and Avinash Pandhare,"*Devnagiri Character Recognition Using Neural Networks*", 2277-3754-7156/$31.00 ©20l3 IEEE, July 2013

[8] Divakar Yadav, Sonia Sánchez-Cuadrado and Jorge Morato,"*Optical Character Recognition for Hindi Language Using a Neural-network Approach*", J Inf Process Syst, Vol.9, No.1, March 2013

[9] Gayathri P and Sonal Ayyappan," *Off-line Handwritten character Recognition using Hidden Markov Model*", 978-1-4799-3080-7114/$31.00 ©20l4 IEEE, 2014

[10] Prof. Mukund R. Joshi and Miss. Vrushali V. Sabale," *Offline character recognition for printed text in Devanagari using Neural Network and Genetic Algorithm*", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 5, May 2015