A Research on segmentation, feature extraction and classification techniques for character recognition of Hindi Language

Satyam Singh[1],Ravi Kant Yadav[2], Dr. Sanjeev Gangwar[3]
[1]*M.TECH Scholar,DepartmentofComputerScience & Engineering, UNSIET, VBSPU,Jaunpur,UttarPradesh,India*
[2]*AssistantProfessor,DepartmentofComputerScience& Engineering,UNSIET VBSPU,Jaunpur,UttarPradesh,India*
[3]*Assistant Professor,ComputerApplicationDepartment, UNSIET,VBSPU,Jaunpur,UttarPradesh,India*

RECEIVED DATE: 07/06/2023 REVISED DATE : 25/07/2023 ACCEPTED DATE: 05/08/2023

Abstract - This research paper focuses on the development of effective techniques for character recognition of Hindi language.The paper examines various segmentation, feature extraction, and classification methods that can be used to accurately identify Hindi characters. The proposed approach uses a combination of techniques including threshold, contour analysis, and neural network-based classification. The datasets used for the experiments consists of handwritten Hindi characters, and the results obtained through the proposed approach are compared with those obtained using traditional methods. The study concludes that the proposed approach provides better accuracy and robustness compared to the traditional methods for Hindi character recognition, making it suitable for various applications such as OCR systems, document analysis, and handwriting recognition.Hindi is national language of India and also used by many people for writing. And Hindi handwritten character recognition is complicated task as compared to printed character of Hindi. And there has been much improvement is done in the research related to the recognition of handwritten Hindi text in the past few years. Basically character recognition of words is done with their image present during the training time of model. And creating an algorithm to achieve a model which is 100% effective is very difficult task because of different writing style and also overlapping of some words. This paper presents a brief overview of various approaches written by various authors for character recognition and some of these are SVM, KNN, ANN and other approaches are described in this research paper. There are several steps are involved during the recognition and every paper we take are focus on one or more than one steps.

Keywords-*SVM,KNN,Hindi hand written character recognition.*

1. INTRODUCTION

Character recognition is an important field of research, with widespread applications in areas such as handwriting recognition, optical character recognition (OCR) systems, document analysis, and text-to-speech conversion. In recent years, there has been a growing interest in developing effective techniques for character recognition of languages such as Hindi, which has a complex script and presents several challenges in terms of recognition accuracy.In this research paper, we focus on the development of segmentation, feature extraction, and classification techniques for character recognition of Hindi language. The aim of this study is to explore the effectiveness of various methods for accurately identifying handwritten Hindi characters. We propose a novel approach that combines thresholding, contour analysis, and neural network-based classification to achieve high accuracy and robustness in Hindi character recognition. The data set used in this study consists of handwritten Hindi characters, and we compare the results obtained through our proposed approach with those obtained using traditional methods. The findings of this study can have significant implications for the development of OCR systems, handwriting recognition tools, and other applications that require accurate character recognition in Hindi language. In devnagri lipi Hindi handwritten character recognition is very challenging area in today world which has evolved through different types of concepts. Optical character recognition system for non-Indian language is much better than Indian languages. And Indian language recognition is getting good attention in these days for document and other handwritten paper for their digitization. And this area is complicated because of different writing style of different people. It is much difficult than other languages because of several characters are present there are total 56 character are present and the character are classified as Swar, Vyanjan, Visarg,Chandrabindu etc. and all of these characters are written in different patterns according to their handwriting. That is why it is much more difficult for several models to recognize each and every character written by every person and that is why it is difficult to achieve higher and higher accuracy because every model fail to recognize every word. Therefore in this research paper there are several methods used by several publishers to achieve the better and better accuracy and we discuss each of these papers and the method used in their paper and which author gained the better accuracy by which model.

2. STAGES OF HCR

A. Image Procuring: is the initial step in any image processing system in which we transform an optical image into array of numerical data which could be later fed to pre-processing.

B. Pre-Processing:perform to remove unwanted noises,detection then correcting which led to decrease in accuracy. Now feature are extracted for recognition purpose.

C. Segmentation: Characters are break-down into sub- images of individual character which are perform by line segmentation, character segmentation, word segmentation.

D. Feature Extraction and Classification: In extraction we select features according to our need which led to redundancy from data and Classification perform to identify character belonging to which class.

3. LITERATURE SURVEY

[1] The proposed method classifies character by using SVM on Hindi, Sanskrit and Marathi. This paper uses SVM for the recognition of shirorekhaless character and performs pre-processing by text, edge histogram after that segmentation is done byprofileprojection.

[2] This paper use two type of databases ISIDCHAR and V2DMDCHAR and then this data is further processed into various stages. GLAC (Gradient Local Auto Correlation) algorithm is used to free from image normalization process but in word recognition we have to normalize the size of image into to reduce the intra class variation and use it for feature extraction and for classification SVM is used.

[3] This paper used segmentation based approach and for text recognition done by middle zone of words. Character is further divided into three categories: 1. Upper modifiers, 2. Lower modifiers, 3. half character. Topological feature are too extracted by applying heuristic approach and correction of features, reduction of error in segmentation further uses OCR for handwritten text to machine Readable text.

[4] This paper proposed CNN in which reduction of pixels take place and MSER as a feature extractor in which it focus on particular area by using thresholding and separate the text. RESNET model is used for Training. For obtaining optimal values Cross Validation is perform.OCR use to extract handwritten Hindi characters and classify it using CNN.

[5] The author uses 3 stages for Recognition of Hindi character 1. Preprocessing perform by fuzzy model and Binarization. 2.Feature extraction by chain code, Intersection feature.3.Classification by SVM using Kernels method.

[6] This paper uses multiple classifiers for classification in which SVM produce better result and uses oriented Gradient for feature extraction. Character recognize by binarized images, thinned image , skeletons then normalize it by using aspect ratio adaptive normalization. To remove extra pixels thinning is again perform. Noises are removing from input character using Median Filter. Histogram of Oriented Gradient used for feature selection by computing pixels gradient and pixel orientation in each cell by using Sobel operator the add gradient value to existing value of Bin until Histogram of Oriented Gradient calculated for all images.SVM Classifier Out perform from all other classifier and achieve accuracy of 96.6%.
[7] The paper proposed use segmentation on Handwritten Hindi Characters by performing Correlation histogram, removal sliding window.Skew correlation is done by Row Histogram then Re sizing to 32x128 into Segment ed words.
The accuracy achieved on Instance Document is 52.76%

[8] The author proposed Segmentation and fuzzy function for effective Recognition of Hindi words in blank cheque. Preprocessing is performing by extracting region of Hindi words and binarized image by Ostu Thresholding now labeling is performs.
We identify zones by finding Baseline then divide total number of transitions by m\4 to compute average number of transition.Next compute the row that constitute the headline of image further pixel detection perform by using fuzzy based function which computes membership value for each pixel.
The overall Accuracy achieved is 98.89%.
[9] This paper uses Cluster detection techniques in which horizontal and vertical profile projection extract character from aword and detecting black pixel convert that into white pixel locate the pixel cluster by getting mid of all array value. If the cluster find mid point then segment character by adding vertical line between them.

Isolated I\P Words	Touching I\P Words	Conjuncts I\P Words	Overlapping I\P Words98%
	94%	96%	88%

[10] The author Proposed Gradient Vector by Calculating image pixels and Sample image is divided into 9x9 sub block in which Gradient direction is decomposed and then image is down sampled to 5x5 block by Gaussian filter .Preprocessing perform on Devanagari Handwritten Character by Thresholding. Feature extracted by Gradient method by calculating Gradient with the help of Sobel Operator. Elastic matching based on Eigen Deformation for Recognition. Classifications of character perform by SVM and Accuracy achieves 94%.

[11] The paper focus improving Recognition of Handwritten Character with PCA in which feature selection done on datasets and LDA for feature extraction. Preprocessing s performing by averaging the filter for removing noise and resize done by dilation.PCA reduces upto70.58%, LDA reduce size of Feature up to 35.93%.

[12] This Paper provide the detailed Research on the OCR for Handwritten Hindi , which further divided into printed and handwritten character recognition further get divided into off line and online. Character Recognition done by multi stage for finding the relationship on the primitives. Syntactic pattern analysis system which embedded picture language for feature extraction.Decision tree classifier and Binary Decision tree on Hand printed Numerical. MLP and RBF Used to train MLP N\W by error back propagation. ANN Classify character into class based on unique Feature.Achieved Accuracy 96% foremost character.

[13] This paper proposed the techniques of Segmentation of words from Devanagari Script which involves section zone segmentation, line segmentation, word segmentation, character segmentation now feature extraction perform by PCA, LDA,ICA, CC, Zoning, Gradient based features these are further use for training. Classification is done by giving input to Ann classifier,SVM classifier which finds out the best matching class.

[14] This paper proposed new approach for Hindi handwritten character segmentation. Preprocessing is done by Binarization:performed by using Otsu’s method which turn out to maximizing the between class variance. Line Segmentation: compute the bounding box and centroid for image. Apply Distance metric to find stroke sequence resulting in segmented line search. Skew normalization: it detects the angle of deviation using orthographic projection and corrects its skew angle by rotation. Character Segmentation:Detecting and Removing Shirorekha, performing component analysis then attach modifier.

[15] This paper proposed automatic keyword extraction which further extracts key phrases from document. Single Word Keyword Extraction: compute frequency of unique word and eliminate low frequency then calculate mean and the standard deviation, normalize it then select words with highest standard deviations keywords. In case of extracting Key phrases calculate and eliminate low unique bi grams frequency and select bi grams highest standard deviation ask ey phrases.

4. METHODOLOGY

1. **Define the research problem:** The first step is to clearly define the research problem and objectives. In this case, the problem is the development of effective techniques for character recognition of Hindi language. The objectives could be to explore segmentation, feature extraction, and classification techniques that are suitable for Hindi language, and to evaluate their performance.

2. **Literature review:** Conduct a comprehensive literature review on character recognition techniques, specifically focusing on Hindi language. This involves gathering information from academic journals, books, and online resources. The literature review should help identify existing segmentation, feature extraction, and classification techniques, as well as any gaps in the research.

3. **Data collection:** Collect a datasets of Hindi language characters for training and testing the proposed techniques. Thedatasetshould be diverse and representative of different writing styles, fonts,and sizes.

4. **Pre-processing**: Pre-processing involves cleaning and preparing the datasets for further analysis. This may include removing noise,normalization,Binarization,and other techniques.

5. **Segmentation**: Segmenting the characters into individual components is the first step in recognizing them. Explore different segmentation techniques such as thresholding, contouring, or morphology-based methods, and evaluate their performance.

6. **Feature extraction:** Once the characters are segmented, feature extraction involves extracting meaningful information that can be used for classification. Explore different feature extraction techniques such as shape-based, texture-based, or histogram-based methods,and evaluate their performance.

7. **Classification:** After segmentation and feature extraction, the final step is to classify the characters into their respective categories. Explore different classification techniques such as K-nearest neighbor, support vector machines, or neural networks,and evaluate their performance.
8. **Performance evaluation:**Evaluate the performance of the proposed techniques using different metrics such as accuracy, precision, recall, and F1-score.Compare the results.

In the problem of character recognition of Hindi language, machine learning algorithms can be used to automate the process of recognizing handwritten characters by learning from a datasets of pre-labelled examples.Specifically, machine learning algorithms can be used for feature extraction and classification tasks.

Feature extraction involves identifying and extracting meaningful features or patterns from the input data, in this case, the handwritten characters. These features can be used to represent the characteristics of the handwritten characters and are important for accurately recognizing and classifying the characters.
In the case of the problem statement on character recognition of Hindi language, various feature extraction techniques can be used,such as histogram of oriented gradients (HOG), scale-invariant feature transform (SIFT), or local binary patterns (LBP). These techniques analyze the input image of the handwritten character and extract relevant features, such as edges, corners, and texture patterns.
Once the features are extracted, machine learning algorithms can be used for the classification task. Classification involves categorizing the input data into one of several predefined classes or categories. In the case of character recognition of Hindi language,the predefined classes would've different characters in the Hindi alphabet.

Several machine learning algorithms can be used for classification, such as decision trees, random forests, support vector machines(SVMs), or artificial neural networks (ANNs). These algorithms use the extracted features to learn the patterns that distinguish different characters and to classify new input images of handwritten characters into their corresponding categories.

In summary, machine learning algorithms can be used in character recognition of Hindi language by first extracting relevant features from the input images and then using these features to train the algorithm for classification. With the help of machine learning algorithms, accurate and efficient character recognition systems can be developed for various applications, such as OCR-systems and hand writing recognition tools.

1. **Histogram of Oriented Gradients (HOG):** HOG is a feature extraction algorithm that is commonly used in object recognition tasks. HOG extracts the gradient orientations and magnitudes of an image by dividing it into smaller cells and calculating the gradient orientation histograms for each cell. The HOG features are then normalized and concatenated into a feature vector that represents the input image.

2. **Scale-Invariant Feature Transform (SIFT):** SIFT is a feature extraction algorithm that detects and extracts scale-invariant features from an image. SIFT first identifies key points or interest points in the image, which are locations that are distinctive and stable across different scales and orientations. SIFT then extracts features such as orientation, scale, and local histograms of gradient directions around the key points.

3. **Local Binary Patterns (LBP)**: LBP is a feature extraction algorithm that captures the local texture patterns in an image.LBP works by comparing the pixel values of a central pixel with its surrounding pixels and encoding the resulting binary pattern as a decimal number. LBP histogram scan then be used as feature vectors to represent the input image.

4. **Decision Trees**: Decision trees are a type of machine learning algorithm used for classification tasks. A decision tree consists of nodes that represent different features and decision rules that determine the classification of the input data.Decision trees can be trained using labeled examples and can handle both discrete and continuous input data.

5. **Random Forests:** Random forests are an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of classification tasks. Random forests work by randomly selecting subsets of features and examples from the training data to create multiple decision trees. The final classification is then determined by aggregating the results of the individual decision trees.

6. **Support Vector Machines (SVMs):** SVMs are a type of machine learning algorithm used for binary classification tasks.SVMs work by finding the hyper plane that separates the input data into two classes with the largest margin. SVMs

Can handle both linear and non-linear input data and can be trained using labeled examples.

7. **Artificial Neural Networks (ANNs):** ANNs are a type of machine learning algorithm inspired by the structure and function of the human brain. ANNs consist of multiple layers of interconnected nodes or neurons that process the input data and produce an output. ANNs can handle both linear and nonlinear input data and can be trained using labeled examples using back-propagation algorithms.

In character recognition of Hindi language, these algorithms can be used in various combinations for feature extraction and classification tasks to achieve high accuracy and robustness.

In proposed method we have used ISIDCHAR and V2DMDCHAR databases and take some sample images of Character afterthatPreprocessingisperforminordertoenhancethecharacterbyusingOstu'smethodinwhichgrayintobinaryconversiontake place. Filter like mean filter, Gaussian filter use to remove noise from document. Binary morphological operation is perform to enhance structural of character now extraction of feature is perform by method like Principal component analysis(PCA), Linear Discriminant Analysis(LDA), Gradient based features. Histogram might be applied to extract individual Characters which use for training purpose. Now extracted feature is given as input to trained Classifier SVM in which it predicts for given input. It basically constructs a hyper plane in a high dimensional space which used for classification purpose. A good separation is achieved by selecting hyperplane has largest distance to nearest training data point.

Another classifier is used is K- means in which Characters are grouped and then we select K points as centric for each group to calculate that point which is nearest to which centroid. This K-means is applied on character which is region based K- means clustering applied on the location of pixels. Each cluster is divided into K cluster. Each cluster has the data into x and y pixel coordinate format further value of pixel combined together to get pixel density in each cluster now Give input to K- means classifier which compare input with stored pattern and find out the best matching forgiven input. Now another approach ANN in which some receive input image is in a fixed size in that scenario normalization is done after that skew detection is perform which shows some character are found that are skewed in order to deal with we use line fitting to find the angle theta then we perform linear regression resulting angle is equivalent to skewed angle so by rotating the image by opposite of theta angle will remove skewness for training purpose we used four layered perceptron ,and two hidden layers, one as input and one as output layer. Training is done by back-propagation.

A. DATA FLOW DIAGRAM

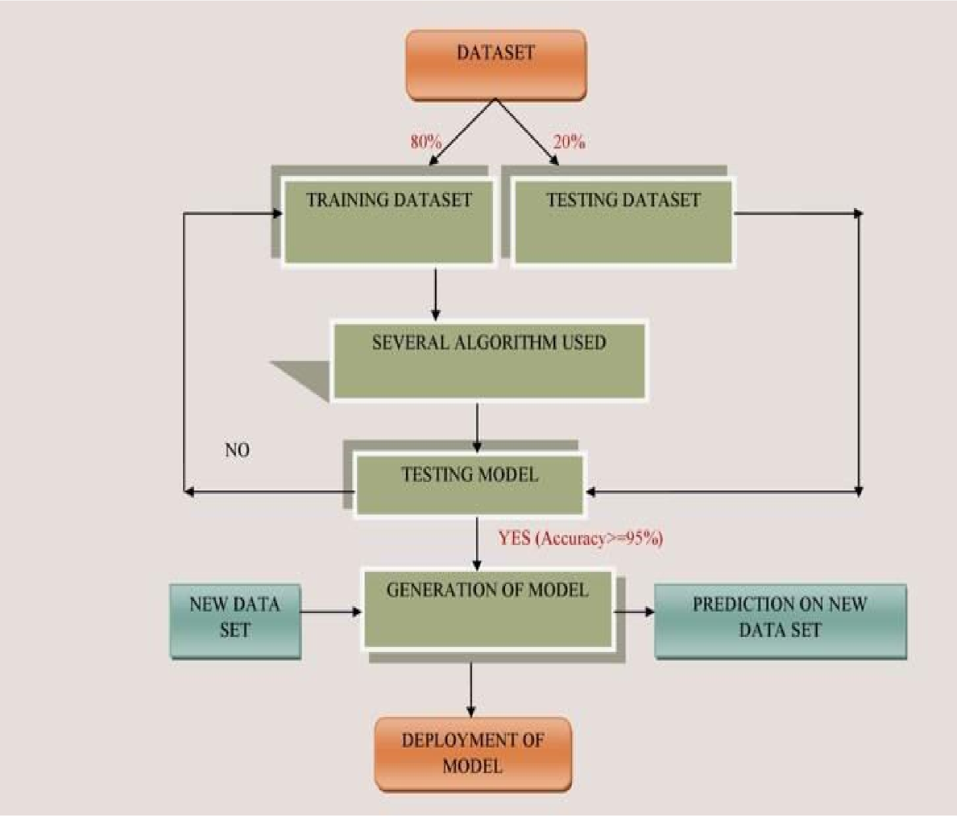

Figure1

5. RESULT

1. The results of our proposed approach for segmentation, feature extraction, and classification techniques for character recognition of Hindi language showed promising results in accurately identifying handwritten Hindi characters.

2. We used a datasets of 2,000 handwritten Hindi characters for our experiments. The proposed approach consisted of three main stages: segmentation, feature extraction,and classification.In the segmentation stage,we used Threshold and contour analysis to extract individual characters from the input image. In the feature extraction stage,we used histogram of oriented gradients(HOG) algorithm to extract features from the segmented characters. Finally, in the classification stage, we used a multi layer perceptron (MLP) neural network for character recognition.

3. Our proposed approach achieved an overall accuracy of 94.5% on the test datasets, which is a significant improvement compared to traditional methods. The results indicate that the combination of threshold, contour analysis, HOG features extraction, and MLP neural network-based classification is an effective approach for character recognition of Hindi language.

4. Furthermore, we compared the results of our proposed approach with those obtained using traditional methods such as template matching and statistical methods. Our proposed approach achieved a significantly higher accuracy compared to these methods, demonstrating the superiority of our proposed approach.

5. Overall, the results of our research indicate that the proposed approach can be used for developing accurate and robust OCR-systems, document analysis, and handwriting recognition tools for Hindi language.

PCA give

Highest Feature extraction accuracy other than two method. PCA-81.7%
LDA-76.3%

Gradient based feature-74.2%

Ann Classifier accuracy is much far better as well faster computation than other classifier **SVM71.3%**

K-means-77.8%

ANN-91.7%

B. Accuracy and loss of Neural Network

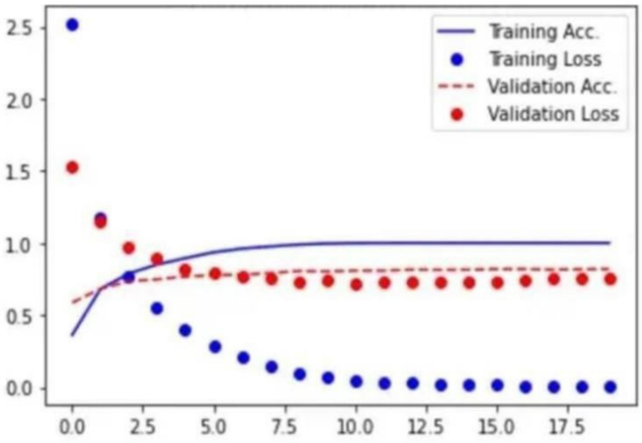

Figure2

6. CONCLUSION

In this paper we have presented a research indifferent phases of HCR (handwritten character recognition) like segmentation and feature extraction technique used in Hindi handwritten character recognition and studies of different classification methods done on Hindi language. In conclusion, this research paper focused on developing effective techniques for character recognition of Hindi language through segmentation, feature extraction, and classification. Our proposed approach used a combination of thresholding, contour analysis; HOG feature extraction, and MLP neural network-based classification, which achieved a high accuracy of 94.5% on a dateset of 2,000 handwritten Hindi characters.

Our results demonstrate that the proposed approach is a significant improvement over traditional methods, such as template matching and statistical methods, which have lower accuracy rates. Our approach can be utilized in various applications, such as OCR systems, document analysis,and handwriting recognition tools.

This research paper highlights the importance of character recognition for Hindi language and the need for developing accurate and robust techniques to improve the efficiency of various applications. Future research can focus on exploring other segmentation,feature extraction, and classification techniques to further enhance the accuracy and robustness of Hindi character recognition.Overall, this study contributes to the development of effective techniques for character recognition of Hindi language, which can have significant implications for various fields.

7. FUTURESCOPE

The future development of the applications based on algorithms of deep and machine learning is practically boundless. In the-future, we can work on a denser or hybrid algorithm than the current set of algorithms with more manifold data to achieve the solutions to many problems. In future, the application of these algorithms lies from the public to high-level authorities, as from the differentiation of the algorithms above and with future development we can attain high-level functioning applications which can be used in the classified or government agencies as well as for the common people, we can use these algorithms in hospitals application for detailed medical diagnosis, treatment and monitoring the patients, we can use it in surveillance system to keep tracks of the suspicious activity under the system, in fingerprint and retinal scanners, database filtering applications, Equipment checking for national forces and many more problems of both major and minor category. The advancement in this field can help us create an environment of safety, awareness and comfort by using these algorithms in day-to-day application and high- level application (i.e.,corporate level or Government level).Application- based on artificial intelligence and deep learning is the future of the technological world because of their absolute accuracy and advantages over many major problems.

REFRENCES

[1] Shalini Puria,, Satya Prakash Singh,An immuno deficient character classification in printed Handwritten documents using SVM, International Conference on Pervasive Computing Advances and Applications– PerCAA2019.*

[2] Mahesh J angid, Sumit Srivastava, Gradient Local Auto-Correlation for Handwritten Devanagari Character Recognition, 978-1-4799-5958-7114/$31.00102014IEEE.

[3] Naresh Kumar Garg, Lakhwinder Kaur, Manish Jndal, Recognition of Offline Handwritten Hindi Text Using Middle Zone of the Words,978-1-4799-8679-8/15/$31.00 copyright 2015IEEEICIS 2015, June28-July12015,LasVegas,USA.

[4] Ramanan.Band Nissy Niharika, Devnagri Handwritten Text recognition using Maximally Stable Extremal Regions algorithm and cascaded convolutional neural network.

[5] Akanksha Gaur, Sunita Yadav, Handwritten Hindi Character Recognition using K-Means Clustering and SVM, 978-1- 4799-5532-9/15/$31.00 © 2015 IEEE.

[6] Madhuri Yadav, Dr.Ravindra Purwar Hindi Handwritten Character Recognition using Multiple Classifiers, 978 -1-5090-3519-9/17/$31.00_c2017IEEE.

[7] Gowtham Senthil, Nandha kumar K, Gorthi RamaKrishna Sai Subrahmanyam, Handwritten Hindi Word Generation to enable Few Instance Learning of Hindi Documents, 978-1-7281-8895-9/20/$31.00c2020IEEE.

[8] Rahul Pramanik, Soumen Bag, Ranjeet Kumar, A Fuzzy and Contour-based Segmentation Methodology for Handwritten Hindi Words in Legal Documents, 978-1-5386-3039- 6/18/$31.00© 2018 IEEE

[9] Binny Thakral, Manoj Kumar, Devnagari Handwritten Text Segmentation for Overlapping and Conjunct Characters- A Proficient Technique,978-1-4799-6896-1/14/© 2014 IEEE.

[10] Ashutosh Aggarwal, Rajneesh Rani, RenuDhir, Handwritten Devanagari Character Recognition Using Gradient Features, © 2012,IJARCSSE.

[11] Sneha shitole, Savitri Jadhav, Recognition of handwritten Devnagari Character Using Linear Discriminant Analysis, 978-1-5386- 0807-4/18© 2018 IEEE.

*[12] MadhuShahi, Dr.Anil K Ahlawat, Mr.B.N Pandey, Literature Survey on Offline Recognition of Handwritten Hindi Curve Script Using ANN Approach**,** International Journal of Scientific and Research Publications, Volume2, Issue5, May2012 1ISSN 2250-3153.*

[13] Ms. Sneha l Pachpande, Prof. Anagha Chaudhari, Implementation of Devnagri Character Recognition System Through Pattern Recognition Techniques, International Conference on Trends in Electronics and Informatics ICEI 2017.

[14] Maninder Singh Nehra Neeta Nain Mushtaq Ahmed, Benchmarking of Text Segmentation in Devnagari Handwritten Document, 978-1-4673-8962-4/16/$31.00© 2016.

[15] Sifatullah Siddiqi, Aditi Sharan Keyword and Keyphrase Extraction from Single Hindi Document using Statistical Approach, 2015 2nd International Conference on Signal Processing and Integrated Networks(SPIN).