International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 02 | Feb 2021 www.irjet.net p-ISSN: 2395-0072

SVM Classifier based Handwritten Character Recognition

Sarvjeet[1], Naveen Dhillon[2]

[1]M.Tech Scholar in R.I.E.T, Phagwara
[2]Principal in R.I.E.T, Phagwara

Abstract : *Developing an android application for character recognition to read the text from an image is a big area of research. Nowadays, there is a trend of storing information from the handwritten documents for future use. The translated machine encoded text can be easily edited, searched and can be processed in many other ways according to requirements. Character recognition systems translate such scanned images of printed, typewritten or handwritten documents into machine encoded text. The method to transform handwritten data into electronic format is Optical Character Recognition. It involves several steps including pre-processing, segmentation, feature extraction and post-processing. Picture information is improved with the help of a technique named image pre-processing. The main challenge is to recognize the characters from different styles of handwriting. Thus, a system is designed that recognizes the handwritten data to obtain an editable text. The output of this system depends upon the data that has to be written by the writer. The representation of samples as points in space that are mapped such that the samples of individual categories can be differentiated using a major vector is known as SVM model. The results show that the proposed system yields good recognition rates which are comparable to that of feature extraction based schemes for handwritten character recognition, high accuracy, precision and recall as compared to existing method.*

Key Words: **Character Detection, Segmentation and SVM.**

1. INTRODUCTION

Handwritten character recognition (HCR) is the process of conversion of handwritten text into machine readable form. The major problem in handwritten character recognition (HCR) system is the variation of the handwriting styles, which can be completely different for different writers. The objective of handwritten character recognition system is to implement user friendly computer assisted character representation that will allow successful extraction of characters from handwritten documents and to digitalize and translate the handwritten text into machine readable text.

In Devanagari script, every character has a horizontal line at the upper portion, termed as Shirorekha or headline. English characters don't have such a trait. This is a distinguishing feature using which English can be extracted from these scripts. In successive handwriting, in the left-to-right direction, shirorekha of a single character connects with the shirorekha of the earlier or subsequent letter of the similar word. Every character and modified shape within a word seems like hanging from the theoretical shirorekha of the word.

Handwritten character Recognition system is divided into two categories

A. On-line character recognition:- It is system in which recognition is performed when characters are Under creation.

- On-line character recognition.

B. Off-line character recognition: It is system in which first handwritten documents are generated, scanned, stored in computer and then they are recognized.

Handwritten Character Recognition System consists of following stages:

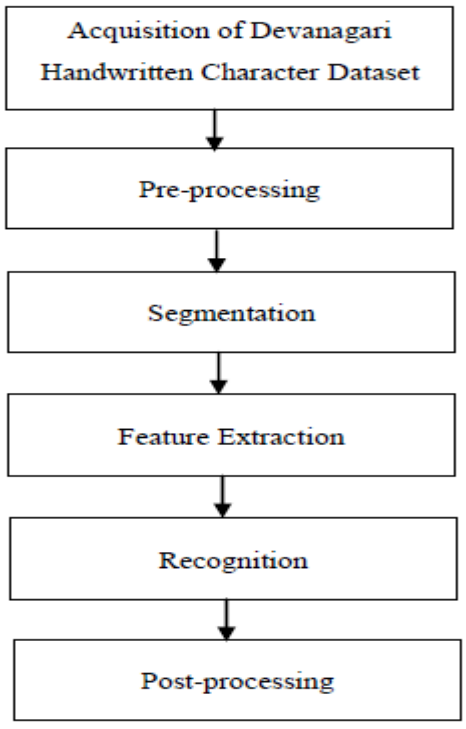

Figure 1: General Process of Devanagari Character Recognition

There are four methods of cursive handwritten word recognition.

- Holistic Approach

1. Holistic Approach: It is method in which entire word is recognized without splitting them by extracting features of entire word.

- Segmentation based Approach

2. Segmentation Based Approach: Characters are segmented from word.

- Recognization based segmentation Approach

3. Recognition based Segmentation Approach: Character classification and segmentation are performed simultaneously by using appropriate learning method.

- Mixed Approach

4. Mixed Approach: This system consist of combination of above methods.

Most organizations use documents to acquire information from customers. These documents are generally Handwritten. Such documents can be forms, checks, etc. For their easier retrieval or information collection documents are transformed and stored in digital formats. Common practice to handle that information is manually filling same data into computer. It would be tiresome and time consuming to handle such documents manually. Hence the requirement of a special Handwritten Character Recognition Software arises which will automatically recognize texts from image of documents. The process of extracting data from the handwritten documents and storing it in electronic formats has made easy by Handwritten Character Recognition (HCR) Software.

2. LITERATURE SURVEY

BrijeshwarDessai, et.al (2019) focused on constructing OPR system based on CNN for Handwritten Devanagari Script so that the characters were recognized in accurate way. Convolutional Neural Networks (CNN) that had more efficiency to classify the image classification than Support Vector Machine (SVM) and Artificial Neural Network (ANNs). Also, the complexity was maximized as a header line was present for every word and uniqueness in writing style. The effectiveness of presented system was enhanced to recognize the Devanagari characters.

H. Bunke, M. Roth and E. G. Schukat-Talamazzini (2018). This paper use holistic method for cursive word recognition. They extract features from the skeleton of word. The feature vector is generated from the edge information of words which includes location of edge relative to four reference line, its curvature, degree of nodes incident to the edge etc. 10-dimensional feature vector is generated .HMM for each letter of alphabets is built and by concatenation of this HMMs, HMM for each dictionary word is built. Limited sized dictionary is used. HMM is trained using Baum-Welch algorithm and reorganization is performed using Viterbi algorithm.

Sneha Shitole, et.al (2018) discussed that the HCR of Devanagari script was the significant field of research in the Pattern recognition. The PCA and LDA had intended for increasing the performance of the recognition system .There were 3 dissimilar feature extraction techniques implemented for extracting the raw attributes. The Linear

Discriminant Analysis (LDA) had mitigated these attributes. The Support Vector Machine (SVM) was deployed for Categorizing the LDA and characters. The outcomes revealed that the LDA had performed more efficiently in Comparison with the PCA.

Sushama Shelke, et.al (2018) introduced the rotation invariant feature extraction methods for handwritten Devanagari characters [24]. This paper made the deployment of various methods namely DCT, DFT and DWT. Additionally, the extraction of their convolved attributes was also performed. It was indicated in the outcomes that attributes of Convolved Wavelet coefficient were proved efficient to extract the rotation invariant attributes in comparison with other transforms. This transform had potential to control the rotation from -30 degrees to +30 degrees in effective way and categorize the character exactly.

D. K. Patel, T. Som and M. K Singh (2012) [7] deals with the handwritten English character recognition using multi resolution technique with Discrete Wavelet Transform (DWT) and Euclidean Distance Metric (EDM). Distances from unknown input pattern vector to all the mean vectors are calculated by EDM. Minimum distance determines the class membership of input pattern vector. EDM gives a recognition accuracy of 90.77%. In case of misclassification, the learning rule through ANN improves the recognition accuracy to 95.38% by comparing scores and then product of generated recognition scores with Euclidean distances has further improves the recognition accuracy to 98.46%.

SushamaShelke, et.al (2016) projected some methods to optimize the accuracy of recognition. First of all, the characters were pre-classified into diverse classes [22]. For this purpose, different structural attributes were utilized. Afterward, optimized feature extraction methods were employed with the intention of extracting the attributes. At last, the neural network (NN) was applied to recognize the character. The deployment of a number of NNs and their analysis had done in this approach. The outcomes exhibited that the recognition rate was enhanced by the means of optimization. The maximum identification rate was acquired using Elman BPNN in comparison with other networks.

3. RESEARCH METHODOLOGY:

Automated Devanagari character recognition is an innovative, prominent and challenging area in today's digitized world, which has evolved through the combination of artificial intelligence, pattern recognition, machine learning and data mining concepts. Optical Character Recognition (OCR) systems of non-Indic languages, such as, English, Chinese, Japanese, Korean, German etc. are already mature as compared to Indic scripts. Because of initial slow progressive growth and much ignorance, Devanagari recognition and classification systems are getting a good deal of attention nowadays. Although many offline Devanagari OCR methods have already been introduced in recent years, yet it is still a big challenge to process its documents due to linguistic based criticalities, large character set, complex conjuncts, and typical geometric structure of character, zone-based form, and use of shirorekha (top line).

Therefore, an efficient Devanagari Character Classification using Support Vector Machine (CC-SVM) method is proposed in this paper, which preprocesses the offline scanned imaged documents, normalizes them, segments them using projection profile, removes top line, obtains Shirorekha-Less (SL) characters, extracts their features, and finally, recognizes the SL characters by using SVM classifier. Such method makes use of a pseudo-thesaurus, which stores a set of pre-defined character classes. The segmented SL characters are matched with the shirorekha based pre-defined characters. If that SL character is found same as the pre-defined character, then it is categorized into that pre-defined character category, otherwise, not.

1. Image preprocessing: After acquiring the scanned text document image preprocessing is performed. At this time, all the variables and counters are initialized, and paths are set to get input images and to store the output variables. All the noise contents and undesired outliers are removed. Its skew is corrected and image is normalized. Further, such image is converted from RGB to gray level image through which the binary image is obtained.

2. Image segmentation: The obtained pre-processed binary image is segmented further to extract the lines, words and characters by using the project profiles. Firstly, the image is segmented through horizontal projection profile, so that all the lines are detected and located in bounding boxes. This is performed because text lines contain high density of black pixels as compared to the gaps existing between the adjacent lines. After such line detection, words are detected and located in bounding boxes by using vertical projection profile. To locate and detect all the characters and their upper/left/right modifiers (orparts), the top lines are removed from image word by word by using horizontal profiling. For this, each top line is first located and then all pixels are made zeroed because it contains the highest density of black pixels in a word.

3. Feature extraction and classification: During this step, the features of all SL character images are extracted. For this, the values are represented in the (rows ×columns) matrix, where the number of rows is equal to number of columns. A character geometrical structure consumes a number of black pixels, which are depicted by its corresponding values, and then their geometric based features are extracted. These extracted features of SL characters and SL modified characters are used to train SVM classifier. Here a pseudo thesaurus is used, which is used to store the pre-defined character classes. Finally, the SL characters and SL characters with 2/more components are matched against these categories.

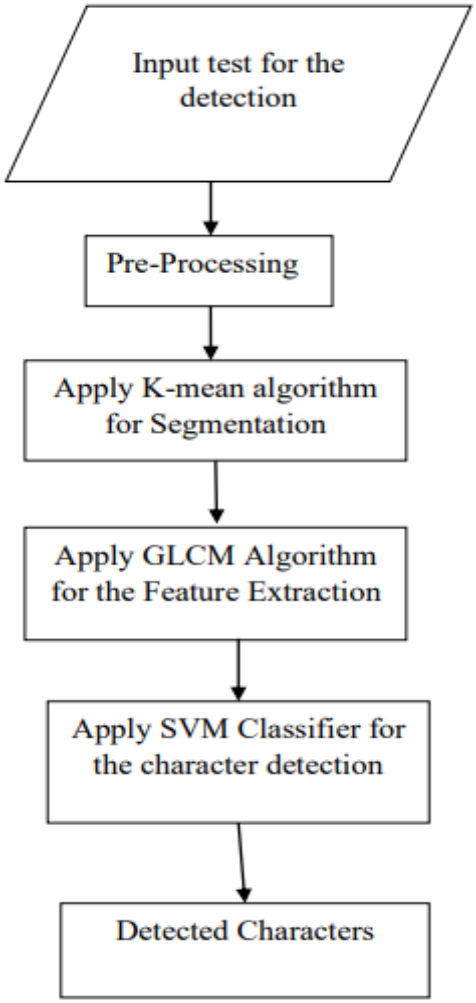

Figure 2: Flow chart of proposed work.

4. EXPERIMENTAL RESULTS

For this implementation, the scanned version of printed documents has been taken, which were collected from different web sites of Government portals, news articles, and traditional shlokas. The handwritten documents are written

by 2 writers in all 3 languages. Total 60 documents are under considered, where 60% documents are used in training and 40% were used during testing. These documents have been evenly distributed among all six document types, such as printed Hindi, printed Sanskrit, printed Marathi, handwritten Hindi, handwritten Sanskrit and handwritten Marathi. Further, this system is also designed to process those images, which have colored background, colored text, multi-colored background and text, bold text, and 12 – 18 pt. font sizes (printed). Some rules have been followed for handwritten documents, and they are given as, Shiro Rekha must be approximately straight, no character or Shiro Rekha overlapping must be there, and characters must not touch each other except the language writing requirements.

The proposed CC-SVM model has been designed for SL character recognition and classification from Hindi, Sanskrit and Marathi imaged documents. The strength of this model lies in its ability to classify the extracted simple SL characters and SL characters along with their upper/right/left components into the pre-defined Shiro Rekha based character categories. It has been found that most of the existing algorithms worked upon Shiro Rekha based characters instead of using SL characters. Second enhancement is that the proposed algorithm extracts complex words and characters from the scanned text imaged documents only, which increases the scope of using diverse set of printed and handwritten imaged documents, whereas it is found that most of the existing algorithms have worked upon the simple character images 43 and their recognition only rather than using the text imaged documents as input. Thirdly, the proposed system works on both character recognition as well as classification, another enhancement on existing technologies.

Document type	SL Classification Accuracy (%)	
	Printed	Handwritten
Hindi	100%	99.23%
Sanskrit	98.77%	97.22%
Marathi	99.86%	98.61%

Table 1: SL classification accuracy results for printed and handwritten Hindi, Sanskrit and Marathi documents

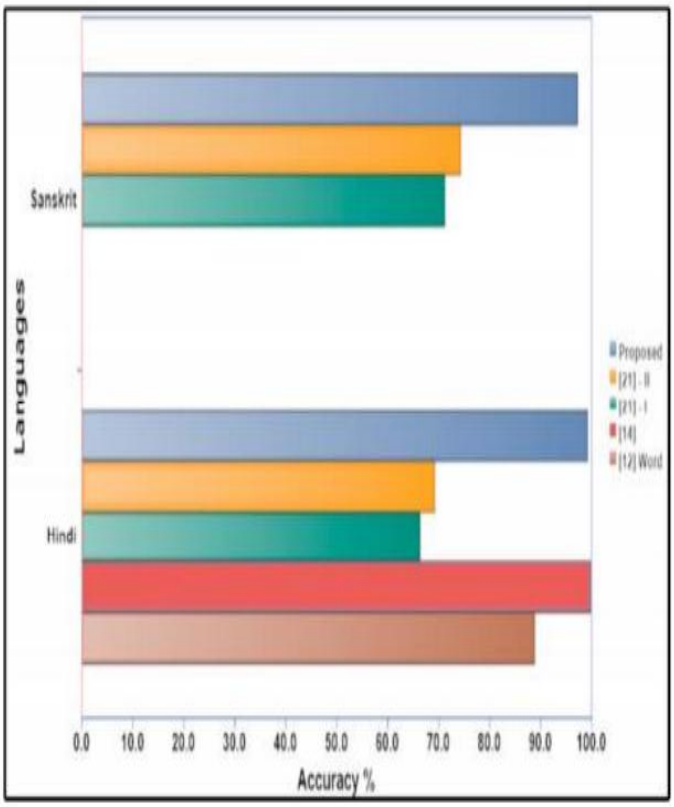

Figure 3: Result comparison of proposed method.

5. CONCLUSIONS

Immense work and research has been done in the handwritten separate character recognition. But so far 100% accuracy is not achieved which gives scope of further work in this direction. Separate characters give good accuracy but word recognition is affected by different writing style. Holistic method eliminates the complicate segmentation but they use limited vocabulary. Segmentation based method due to its complexity acquire less accuracy. Good accuracy is observed in the classifier where scope of words is limited to fix numbers as it has to deal with limited number of variation.

In future, the proposed system will be improved for a large set of printed and handwritten document images.

- Recognition and classification of image words.
- Improving the system for multi-fonts and italics text.
- Extending the work for imaged document classification.
- Making the proposed system generic for the inclusion of other Indic and non-Indic scripts.

REFERENCES

[1] Vasu Negi, Suman Mann, Vivek Chauhan, "Devanagari Character Recognition Using Artificial Neural Network", 2017, International Journal of Engineering and Technology (IJET), vol 9, no. 3.

[2] Mahesh Jangid and Sumit Srivastava, "Handwritten Devanagari Character Recognition Using Layer-Wise Training of Deep Convolutional Neural Networks and Adaptive Gradient Methods", 2018, Journal of Imaging.

[3] Vikas J Dongre, Vijay H Mankar, "A Review of Research on Devnagari Character Recognition", 2010, International Journal of Computer Applications, Volume 12, No. 2, pp. 0975 – 8887.

[4] D. K. Patel, T. Som and M. K Singh," Improving the Recognition of Handwritten Characters using Neural Network through Multiresolution Technique and Euclidean Distance Metric", International Journal of Computer Applications (0975 – 8887) Volume 45– No.6 May 2012.

[5] Richa Patil, VarunakshiBhojane, "Character recognition of Devanagari characters using Artificial Neural Network", 2015, International Journal of Computational Engineering Research (IJCER), Volume 5, Issue 12.

[6] Ankita Srivastav, Neha Sahu, "Segmentation of Devanagari Handwritten Characters", 2016, International Journalof Computer Applications (0975 – 8887) Volume 142, No. 14.

[7] BrijeshwarDessai, Amit Patil, "A Deep Learning Approach for Optical Character Recognition of Handwritten Devanagari Script", 2019, 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT).

[8] Vivek Kumar Verma, Pradeep Kumar Tiwari, "Removal of Obstacles in Devanagari Script for Efficient Optical Character Recognition", 2015, International Conference on Computational Intelligence and Communication Networks(CICN).

[9]Amit Choudhary, Rahul Rishi and Savita Ahlawat, "Off-Line Handwritten Character Recognition using Features Extracted from Binarization Technique", 2212-6716 © 2013 American Applied Science Research Institute doi:10.1016/j.aasri.2013.10.045.

[10] Anshul Gupta, Manisha Srivastava and Chitralekha Mahanta," Offline Handwritten Character Recognition International Conference on Computer Applications and Industrial Electronics (ICCAIE), 2011.