TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 5494 - 5500

Analysis on Devanagari Text Generation by Using Machine Learning Techniques

[1]Mr.Vajid Khan, [1]Dr.Yogesh Kumar Sharma

[1]Research Scholar, [2]Professor, Research Coordinator JJTU Rajasthan

[1-2]Deparment of Computer science and engineering,

[1]Research Scholar, Shri JJTU, Rajasthan, Professors, [2]Head of Department, Research Coordinator Shri JJTU, Rajasthan,

Email id: kvajid12@gmail.com, dr.yogeshkumar@yahoo.in,

Article Info
Volume 82
Page Number: 5494 - 5500
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 27 January 2020*

Abstract:
Many Authors are developed different methods for recognition of the Devanagari script. The existing method is processing to recognize the Devanagari script which is discussed with notable performances. Generally, the recognition process mainly consists of three steps pre-processing, extraction of feature and finally classification. For character recognition different algorithm were being developed with different advances which include neural network algorithm (NNA), pattern matching algorithm (PMA), structural algorithm (SA), support vector machine algorithm (SVMA), statistical algorithm, hidden Markov model (HMM) and template matching algorithm(TMA). In template matching algorithm only the typewritten characters can be recognized. But the other algorithm like neural network algorithm (NNA), structural algorithm (SA), and support vector machine (SVM) can recognize both handwritten and typewritten Every algorithm contains both advantages and disadvantages. The main motivation of this thesis work is to overcome the above mention the drawbacks.

Keywords: *: Handwriting recognition, CNN-RNN network, Data augmentation, Image pre-processing, Optical Character Recognition, Artificial Intelligence, Deep Neural Network, Deep Learning, Devanagari Script*

I. INTRODUCTION

In India, many different languages were scheduled. These languages were written using different types of scripts. Generally the scripts used for writing the language is classified into Indic and non-Indic script. Within Indic script the Devanagari script is the basic script and is used for writing many languages in India. Some of the languages are mentioned as follows, which includes Assamese, Bengali, Bodo, Dogari, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, and Urdu. In addition to these languages, hundreds of other languages are also used in India. These languages were used to describe many documents. These documents must be converted into

Readable format. The boundless world of Internet has afforded the digital format to numerous authors from diversified domains to express their views and knowledge. These authors' valuable data can be easily written and shared to intended readers in digital format rather than typical paper format. This digital format gives ease of editing, formatting, and quick sharing of the information [226]. To converting these paper format into digital format many process were being carried out. The process that is carried out in these will be described in the upcoming section. And in the following section an overview regarding the work is described below.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 5494 - 5500

The Devanagari script is found to be the basic script among all other scripts and it is used in many languages. Many documents were written using this script. For easy understand of the content by the people many different process were carried out in these documents. In today's world due to advances in technology the paper format document can be converted into digital format through various process. There are many issues present in this script which includes variability in writing style, existence of multiple forms of writing the same character, existence of touching and fused characters, lack of standard benchmarking and ground truth dataset, lack of corpora and complexity of grammatical formation of the sentence. Due to these issues the character recognition and text generation in Devanagari script is found to complex task. Many works were not reported in this because of these challenges and difficulties.

Nowadays many authors were focusing in this area since it is found to be a challenging task. Different techniques were used for processing these documents. The process involved in this was described as follows. The document will be given as the scanned image. The characters present in the scanned image must be first segmented by means of different methods. Following that the recognition process is carried out through different process. The character recognition in Devanagari character is more complex due to shape of constituent stroke [227,228]. After the completion of character recognition process the text will be generated at the final process. A brief outline about the Devanagari script is discussed in the following section.

Text generation is one of the subfields of natural language processing. It was knowledge leverages on the artificial intelligence and computational linguistics for the natural language texts generated by naturally, certain communicative requirements can have satisfied by these texts. The combination of practical and different commercial applications is called text generation, in which manuscripts, reading forms and their archival etc. Like as a less uses only interact the keyboard facilitates system. As well as the text is moreover hand-written or printed which terms are directly transferred to the machine. Mainly the visually handicapped getting more benefits by this text generation when interfaced with a voice synthesizer.

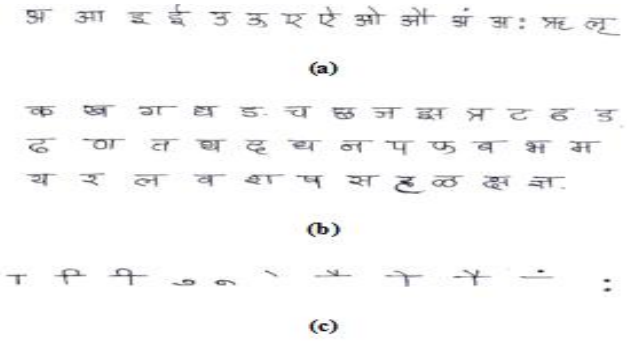

Figure: 1 Devanagari Text Generation

The building changes of text recognition system can combine the performance of human competition as well as providing a strong motivation to the field researches. The creation of Devanagari script is nature and symbols and character composition is being in the two-dimensional word that can make the difference from the ideographic and Roman scripts. In-text generation script involves (a) vowels, (b) consonant and (c) modifiers. The Devanagari text generation is given in figure 1.

Handwriting Character Recognition (HCR) issues are frequently expressed as symbol or isolated character classification task which followed through a stage of post-classification (that is consisting modules such error correction, UNICODE generation, etc.) to generate the textual illustration, for main of the Indian scripts. Some approaches are disposed to the failures due to,

1. In designing difficulties, depending on the word-to-symbol segmentation module which can strongly work on the presence of degraded (fused/cut) images.
2. Outputs of the classifiers are converting to a UNICODE valid sequence. And look at two important aspects of word recognition word image to text string conversion and error detection and correction in words represented as UNICODE.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 5494 - 5500

II. LITURATURE SURVEY

2.1 Devanagari Handwritten Recognition

Normally the handwriting recognition system is containing five kinds of parts such as image acquisition, image pre-processing, feature extraction, character recognition and display result. The illustration of the recognition system is shown in figure 2.

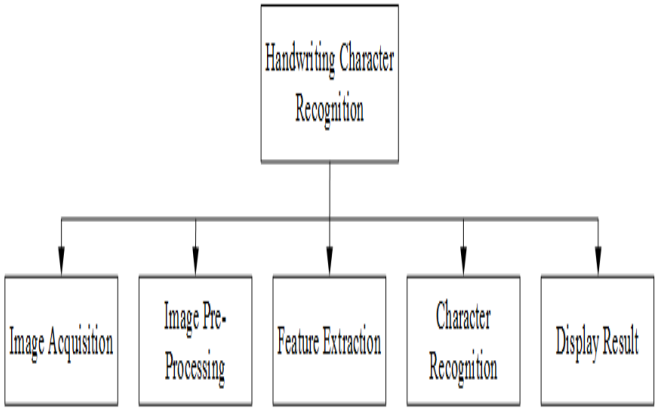

Figure: 2 Handwritten Character Recognition

Image Acquisition: Input of the image is scanned in image acquisition by recognition system obtains

Pre-Processing: It is one of the series operation performances in the input scanner image.

Image Pre-Processing: This kind of techniques is necessary on the binary document, color or grey-level images which consisting graphics or text.

Segmentation: Image sequence of characters is disintegrated into individual character's sub-images on the segmentation stage

Method of Feature Extraction: Features using characters are critical for their classification in the extracted recognition stage. The extraction stage is a very important stage for the enhancement of effective function such as misclassification reduction and recognition rate.

Character Reorganization: Character recognition stage is the decision making part of a recognition system and this stage is using the features extracted in the previous stage.

Finally Output: In figure 3, part a and b are showing the Hindi printed and handwriting samples, part c and d are showing Sanskrit printed and handwriting samples and part e and f are showing Marathi printed and handwriting samples

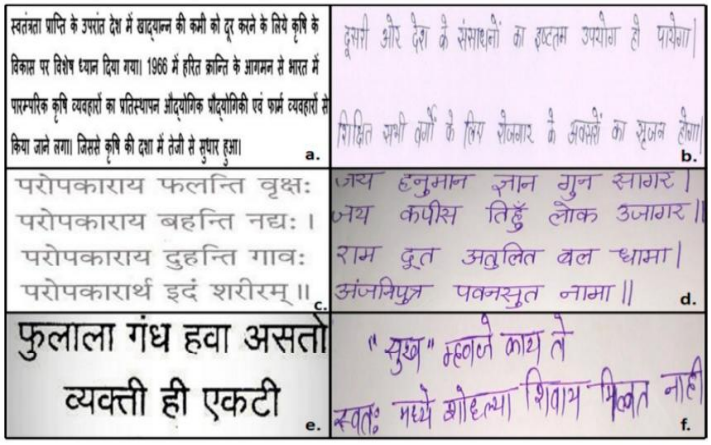

Figure: 3 Sample Document Set for Printed and Handwritten

Devanagari Text Generation Machine Learning Techniques:

Machine learning techniques are mostly helped to the hand-printed Devanagari characters. That recognition issues have attracted more of the work researches in the field pattern recognition, due to the commercial possible application of character recognition techniques and the comfort availability of raw data used for the implementation and testing of different framework general proposition of pattern recognition. The machine learning is one of the subfields of artificial intelligence which concerned with the development and design of techniques and algorithm that terms are allowed to "learn" on the computers [38]. Broadly, four kinds of classifications are used for machine learning.

Devanagari script is designed by curves, semi-circular lines, circular, cross, horizontal and vertical. The Devanagari script writing format is left to right. The attached format of the character is one below the other or forward sequence. Two ways are can be used to do the handwriting character

January - February 2020
ISSN: 0193 - 4120 Page No. 5494 - 5500

recognition such as online and offline. But some of the conversions are occurring on the writing character recognition such as handwriting text into digital text or scanned image into print. In offline character recognition, scanner is used to write the capture optical during successive coordination points in on-line character recognition and they are the function time is made well strokes through consideration of users. The branch of optical character recognition is a handwriting character recognition that can be converted the input handwriting from a paper document into digital text [42]. The classification of the handwriting recognition method is given in figure 4.

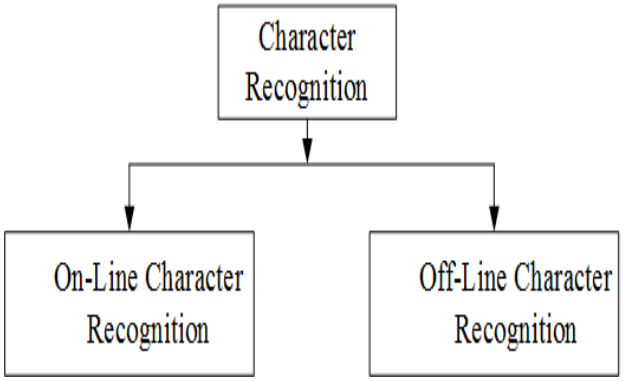

Figure: 4 Types of Handwriting Character Recognition

Off-Line Recognition: Automatic conversion of text is involving in the off-line handwriting recognition, and then the image converted into letter codes that are practical within text-processing and computer applications. The data attained through this method is observed as a static handwriting representation. The off-line handwriting recognition is difficult when compared to other recognition because every people follow various kind of handwriting styles. And, nowadays, handwriting character recognition engines are focused primarily on ICR and machine-printed text for hand-printed text (using only capital letters).

On-Line Recognition: Automatic conversion of text involved in the On-line handwriting recognition, as the written character on a PDA or special digitizer, where a sensor picks up the pen-tip actions and also pen-down or pen-up switching. This data type is called as digital ink and it can be used to observe the digital representation of handwriting [43]. The gained signal is converted into letter codes that are operating within text-processing and computer applications. The elements of on-line handwriting recognition interface characteristically include:

Hidden Markov Model (HMM):

A hidden Markov model (HMM) is based on the offline handwritten Devanagari words. The chain-code histogram directions on the image strips scanned from left to right through a window sliding, which terms are used as the feature vector. HMM continuous density is recognizing as a word image. One of the important roles is HMM plays, in recognition of cursive handwriting. The main reason for overdue which that stochastic models, it can have faced the differentiation in patterns at more efficiently. From there is having more difference in handwriting styles, an HMM is one of the natural choices to the tool of handwriting recognition. Another one most important factor of HMM using is which evades categorical word segmentation and also holistic activities approach through with contextual knowledge and local features

Support Vector Machine (SVM):

Support Vector Machines (SVM) is widely using on the pattern recognition for the classification. The support vector machine is also known as the supervised learning method. In the early nineties, practically first implementation method is SVM which is also known as the most efficient family of algorithms on computationally efficient and machine learning. The support vector machines are mostly used for the learning purpose, for that using hypothesis space of linear functions on a trained, high dimensional feature space with learning algorithm since implementation optimal theory of learning bias and that is derived from the theory of statistical learning. Support Vector Machine acts as a supervised Machine Learning technique.

Convolution Neural Network (CNN):

Deep Convolutional Neural Network (CNN) has given superior results for the shallow traditional networks on more of the recognition tasks. Deep convolution neural network is keeping the distance with the regular approach of character recognition, as

January - February 2020
ISSN: 0193 - 4120 Page No. 5494 - 5500

well as focused on the increment approach of dataset and dropout for the test accuracy enhancement. The nonlinear hidden layers are considered on the deep neural network, so the trainable parameters and the numbers of connections are being larger. As wellbeing harder to train, that is large set of the network examples are required for the overfitting prevention. One set of the deep neural network compared with the smaller parameter group, this method is very easy to train the CNN. Convolution neural network ability to correctly input dataset model can be differing through change the hidden number of the layer in trainable parameters of each layer, as well as correct assumptions made on the natural image [50]. Such a standard forward feed network, they can make the complex model of non-linear relationship within input and output.

Recurrent Neural Network (RNN):

Recurrent neural network (RNNs) is a connectionist model which consisting a self-connection of the hidden layer. RNN provided is a more sophisticated way of dealing with time or sequential data which correlations exemplifies between data points, they have been closed on the sequence. RNN has been applied successfully to the cursive handwritten document recognition, in scripts like Arabic and English. The consistent recurrent neural network (RNN) is protracted for the bidirectional recurrent neural network (BRNN).

III. METHODOLOGY

Handwritten recognition and text generation is currently getting the attention of researchers because of possible applications in assisting technology for blind and visually impaired users, human–robot interaction, automatic data entry for business documents, etc. Here, I have intended to develop Ant Miner Algorithm (AMA) for recognition and text generation of the Devanagari scripts. The proposed Devanagari text generation system using AMA is designed to accept scanned document images which pre-processes, segments, feature extraction and lastly recognition and text generation. Generally the AMA architecture is consists of two phases such as training and testing phases. During the training phase, a known set of text document images are taken which are further processed to get Devanagari characters. When top lines are cleared from words during segmentation process, then independent and isolated Devanagari characters are obtained. In the segmentation process, the Devanagari characters can either be non-modified simple characters or those characters whose modifiers are separated. After that, the segmented character features are extracted for training the AMA algorithm. Initially, the AMA is calculated the minimum distance among the character and its upper/right/left modifier. After that, the recognition and text generation is attained with the help of AMA algorithm. During testing phase, the performance of the AMA algorithm is tested with an unknown set of document images. The proposed method will be compared with the existing methods of Artificial Neural Network (ANN) and Deep Learning (DL).

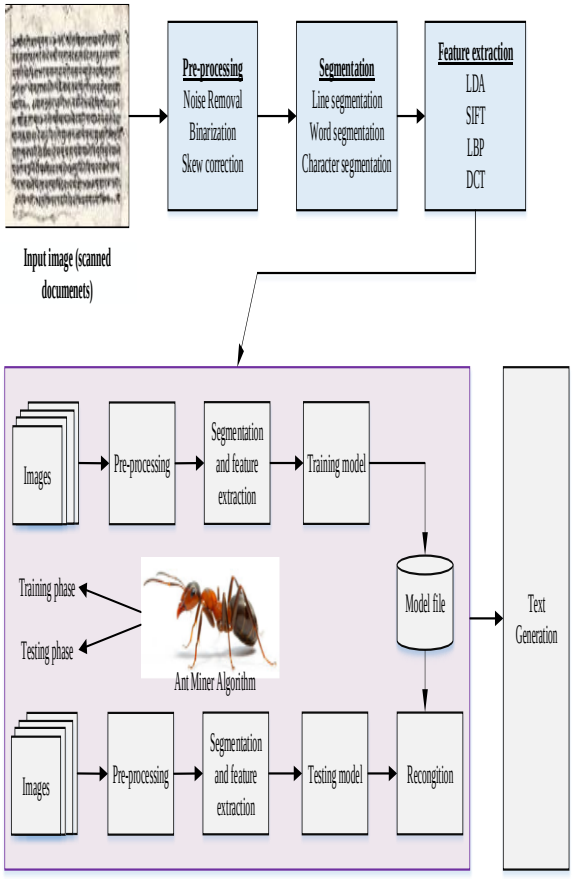

Figure 5: Architecture of the proposed method

Process of the proposed methodology

January - February 2020
ISSN: 0193 - 4120 Page No. 5494 - 5500

Step 1: Input image of Devanagari Script handwritten and scanned images are collected from the open source system.
Step 2: Pre-processing Stage: Noise removal, Banalization and Skew correction
Step 3: Segmentation: Line segmentation, word segmentation and character segmentation
Step 4: Feature extraction: Linear discriminant Analysis (LDA), Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBP) and Discrete Cosine Transform (DCT).
Step 5: AMA algorithm is working based on the two phases such as training phase and testing phase. In training phase, the features are training after that testing phase which can be tested.
Step 6: Finally, Devanagari handwritten recognized and text generation is achieved with the help of the AMA algorithm.
Simulation Tool: Python

IV. CONCLUSION

The basics of handwriting character recognition and its various stages. In this Paper, explained discussed the text recognition, handwriting recognition, printed recognition of Devanagari script. Methods used for classification such as the Hidden Markov Model, support vector machine, Convolution neural network and Recurrent neural network are described. The various assumptions, objectives and major contributions of this research work have also been described

V.REFERENCESS

1. Ghosh, Rajib, Saurav Shanu, Sugandha Ranjan, and Khusboo Kumari. "An approach based on classifier combination for online handwritten text and non-text classification in Devanagari script." Sādhanā, Vol. 44, No. 8, pp: 178, 2019.

2. Singh, Pawan Kumar, Supratim Das, Ram Sarkar, and Mita Nasipuri. "Script invariant handwritten digit recognition using a simple feature descriptor." International Journal of Computational Vision and Robotics, Vol. 8, No. 5, pp: 543-560, 2018.
3. Deore, Shalaka Prasad, and Albert Pravin. "On-Line Devanagari Handwritten Character Recognition Using Moments Features." In International Conference on Recent Trends in Image Processing and Pattern Recognition, Vol. 1, No. 1, pp. 37-48, 2018.
4. Bhalerao, Milind, Sanjiv Bonde, Abhijeet Nandedkar, and Sushma Pilawan. "Combined classifier approach for offline handwritten Devanagari character recognition using multiple features." In Computational Vision and Bio Inspired Computing, Vol. 1, No. 1, pp. 45-54, 2018.
5. Kaur, Simerpreet. "Devanagari and Gurmukhi Handwritten Character Generation using Generative Adversarial Networks." PhD diss., Vol. 1, No. 1, 2018.
6. Pawlewski, Pawel. "Script language to describe agent's behaviors." In International Conference on Practical Applications of Agents and Multi-Agent Systems, Vol. 1, No. 1, pp. 137-148, 2018.
7. Kumar, Munish, M. K. Jindal, R. K. Sharma, and Simpel Rani Jindal. "Character and numeral recognition for non-Indic and Indic scripts: a survey." Artificial Intelligence Review, Vol. 1, No. 1, pp: 1-27, 2018.
8. Alghamdi, Mansoor, and William Teahan. "Printed Arabic Script Recognition: A Survey." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, Vol. 9, No. 9, pp: 415-428, 2018.
9. Kiessling, Benjamin, Daniel Stökl Ben Ezra, and Matthew Thomas Miller. "BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts." arXiv preprint arXiv, Vol. 1, No. 1, pp:1907.04041, 2019.
10. Lam, Kai-Chee, Lay-Hoon Ang, Wee-Ling Kuan, and Foo-Terng Hoe. "Character recognition through wild association: An alternative in learning Chinese script for beginners." Issues in Language Studies, Vol. 7, No. 1, pp: 1-11, 2018.
11. Framework for Privacy Preserving Classification in Data Mining ,DYK Sharma, GM Sharif, Journal of Emerging Technologies and Innovative Research 5 (9), 178-183
12. Performance Evaluation of Delay Tolerant Networks Routing Protocols Under Varying Time of Live, VK Samyal, DYK Sharma International Journal of Advance Research in Computer Science (IJARCS) 8
13. Li-Fi the Most Recent Innovation in Wireless Communication S Saini, DYK Sharma International Journal of Advanced Research in Computer Science and Software.

January - February 2020
ISSN: 0193 - 4120 Page No. 5494 - 5500

14. Web Page Classification on News Feeds Using Hybrid Technique for Extraction AD Patel, DYK Sharma Information & Communication Technology for intelligent system 107 (6), 399-405.
15. Impact of buffer size on different drop policies (DLR, MOFO and E-Drop) for MaxProp Routing Protocol in DTN VK Samyal, DYK Sharma International Journal of Research in Applied Science & Engineering
16. A Pragmatic Evaluation of Stress and Performance Testing Technologies for Web Based Applications P sonkari, DYK Sharma IEEE, Amity International Conference on Artificial Intelligence (AICAI), 399-403
17. Impact of Network Load and node Mobility on the performance of Proactive, Reactive and Hybrid routing protocols of MANET V Singla, YK Sharma International Journal of Advanced Research in Computer Science 8
18. Analysis of Selfish Node Behavior in Delay Tolerant Networks Routing Protocols VK Samyal, DYK Sharma International Journal of Innovative Research in Science and Engineering
19. SIGNIFICANCE STUDY OF USER WEB ACCESS RECORDS MINING FOR BUSINESS INTELLIGENCE, AD Vyas, YK Sharma Indian Journal of Applied Research
20. A COMPREHENSIVE STUDY ON CLASSIFICATION OF AUTOMATED CATEGORIZATION OF WEB SITES: A PROPOSED METHODOLOGY DA Vyas, DYK Sharma INDIAN JOURNAL OF APPLIED RESEARCH (IJAR)
21. Using Open CV for Machine Learning in Real Time Computer Vision and Image ProcessingP S, YK SharmaInternational Journal of Recent Technology and Engineering