2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418028

Character Recognition System for Devanagari Script Using Machine Learning Approach

Shilpa Mangesh Pande
Associate Professor, Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, India and affiliated to Visvesvaraya Technological University, Belagavi
shilpa.p@cmrit.ac.in

Bineet Kumar Jha
Assistant Professor, Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, India and affiliated to Visvesvaraya Technological University, Belagavi
yoursbineetjha@gmail.com

***Abstract* — It is a very difficult task to manually process the handwritten documents due to varieties of handwritten scripts and lack of associated language dictionary to interpret documents. Most of the large companies as well as small-scale industries want to automate the process of script recognition. The big challenge is to make machines recognize the hand-printed scripts. Humans can recognize handwritten or hand-printed words after gaining knowledge of a specific language. In the same way, machines should be trained to recognize the handwritten scripts. This process of transferring human knowledge to computers should be automated. The proposed research work attempts to automate the character recognition system for Devanagari script using various machine learning classifiers like Decision Tree classifier, Nearest Centroid classifier, K Nearest Neighbors classifier, Extra Trees classifiers and Random Forest classifier. The performance of all the classifiers is evaluated using accuracy parameter as success criteria. The Extra Trees classifiers and Random Forest classifier is proved to better than other classifiers with 78% and 77% of accuracy respectively. The robustness to picture quality, writing style, font size is the novelty of the OCR system which makes it ideal to use.**

Keywords— Character recognition system, Devanagari script, K Nearest Neighbors classifier, Decision Tree classifier, Extra Trees classifiers, Random Forest classifier

I. Introduction

Machines are used to ease the human beings works and expand their capabilities by amplifying cognitive strengths. Nowadays machines are performing human's daily tasks such as writing, reading, reminding, suggesting alternatives. Humans are training machines to get their own works done by them. Human skills should be taught to computers so they can perform the assigned jobs very well. This requirement created multiple opportunities in many research fields such as recognition of optical characteristics, neural networks, machine learning, artificial intelligence, robotics pattern recognition, and many more. Machines are being used to automate the works in most of the IT companies as well as in the small-scale industries, educational institutions, manufacturing hubs, banking sectors, various government offices, and private industry. This automation helped to speed up and get the consistent quality of the work.

In the early nineties, a lot of handwritten data was processed manually [1] which includes addresses written on envelopes, bank cheques, phone bills, electricity bills, and letters. Most of the hand-printed data is available in the public, government sector and insurance sector. It is very difficult to process handwritten data manually. Manual processing of all the data is a tedious job and needs a lot of patience and hard work. This may lead to a lot of human errors. The optical character recognition (OCR) converts the hand-printed data into an electronic form so the machines can recognize it [2]. It is an emerging field. The key advantage of this digitization technique is that scanned documents can be edited which is not possible with handwritten scripts. The OCR is very much like a handheld scanner that is used to read OMR exam sheets or a barcode. The OCR reads the handwritten data and can classify it into various classes. The main two classes to classify OCR are online and offline. Tablets, an electronic pad use an online character recognition system whereas scanned documents, handwritten scripts use offline character recognition systems [3]. This paper focuses on offline character recognition. The preprocessing, feature extraction and classification are the main steps in the basic OCR system. The character segmentation, skew detection and correction, thinning, binarization are the methods involved in the preprocessing the data. The character segmentation is a mechanism which separates a sequence of characters into the individual characters. Based on segmented characters, the feature set is prepared and then the character is classified into the specific class. The character segmentation plays an important role in developing OCR system and its performance. Cursive handwriting is difficult to segment because of a wide variety of writing styles of various authors, poor quality of documents [3].

In this paper, a character recognition system is developed for Devanagari script. The Devanagari script is the origin of more than a hundred languages spoken in India like Sanskrit, Hindi, Marathi, and Maithili. The Devanagari script consists of 47 main alphabets, 33 consonants, 14 vowels and 10 digits. Apart from that, the alphabets get changed when added with the vowel is added to the consonants. The machine learning approach is used to recognize the characters and classify them. The system is robust for the picture quality, font size, writing style which makes it ideal to use.

The rest of the paper is organized as follows. Section II discusses the background work of researchers in the field of character recognition. Section III proposes a model for the character recognition system. Section IV discusses the results and the observations, and Section V is the conclusion of the paper.

II. Related Work

The research work is carried out on the Marathi script for document analysis and classification digitally. The recognition of handwritten words is a tough job because of multiple aspects. For character recognition, the vertical projection is used to fragment the words and for line segmentation and horizontal histogram approach is customized. Using the bounding box method, the words are split up which provided segmentation accuracy [4].

In [5] this paper, with the help of various knowledge sources, Devanagari script recognition is highlighted. These information sources are used in the script identification in a hierarchical way. Most of the researchers focused on both printed manuscripts as well as on hand-written ones. For segmentation and identification of image documents, multiple characteristics like curves, height, colors are considered [6,7]. Mumford-shah, profile projection and Hough transform techniques have been implemented for many languages [8,9,10]. The revamped segmentation approach is used for Devanagari text line and word segmentation [11]. Several mixed script images, handwritten data, noisy data and low-resolution parameters are used in these systems.

Statistical and structural features are utilized for the recognition of characters. The Multi-layer Perceptron classifier provided effective precision [12]. Different methods for handwritten and printed scripts have been surveyed [13] and developed the recognition system using k-NN classifier using curvelet transform [14]. As a function of handwritten characters, regular expressions are used. The proposed approach is used in many languages such as Arabic, English and Chinese [15]. The classifier used minimum edit distance to calculate the accuracy. Multiple thinning methodologies are discussed [16, 17, 18]. The Global Alignment Algorithm is implemented to develop an online character recognition system. To get the most profitable decision multiple hypothesis tree is used [19]. The optical character recognition system is designed using Convolution Neural Networks (CNN) for the Sanskrit manuscript. An image segmentation algorithm is used to compute pixel intensities to recognize the alphabets in the images [21]. The paper illustrates the system to identify the characters in the Konkani scripts using Artificial Neural Networks (ANN) and then converts the text into speech. The framework is built to convert text to speech using a speech synthesizer and recorded the accuracy of 61% [22].

The Optical Character Recognition System is proposed for visually challenged. The framework is designed using Raspberry pi, image segmentation algorithms and a Text-to-Speech synthesizer (TTS) [23]. Marathi script to English script translator is developed using sentence tokenization. A bilingual dictionary is used to find meaning of each English word which is separated as a token [24]. Support Vector Machine Classifier is implemented to identify the handwritten Hindi scripts using self generated database. The hand printed documents are segmented at three levels at line word and character. Morphological operations and various things methods are suggested to recognize the text [25].

The Capsules Network (Caps-Net) based multitask learning architecture is used for text classification and it also minimizes inference in multitask learning [26]. The Capsules Network (Caps-Net) along with the font style classification algorithm is used to classify times new roman, Algerian, Arial black font style in English alphabets. The proposed work is compared with various other algorithms like K Nearest Neighbor (KNN), Decision Tree (DT) and Naïve Bayes (NB). The Caps-Net classified the images with better accuracy than other algorithms [27].

III. Models and Methodology

A. Database Creation

There is no standard way to create the database for the Devanagari script for the offline character recognition system. The proposed technique here initially creates the database of Devanagari script alphabets. The Devanagari alphabets are written on paper and then scanned as images. First Authors et al [14] have collected the data from other authors and the same data is used for character recognition. A flat bed scanner at 300dpi is used to create the database [20]. In this proposed work, an offline database consists of 43 thousand images (32 × 32 pixels) of 43 characters, consonants ka to gya and digits 0 to 9. To ensure different sizes and orientations one thousand variations are considered for each character. To make the system more efficient feature extraction and character classification techniques are used. The database is created very carefully in such a way that the basic form of the character is not changed. The font size, writing styles and picture quality are the main parameters focused on the robustness of the system. Fig. 1 shows the sample database of the scanned images with various variations. The database consists of all the letters from the Devanagari script with a variation in writing style, font size.

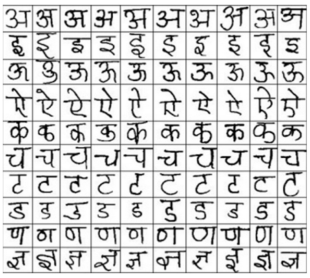

Fig. 1 Sample Scanned Image Dataset

B. *Recognize the character*

Characters can be portrayed in multiple ways such as profiles, gray images, thinned images, skeletons, contours etc. Choosing the right character tends to depend upon the type of classifier as well as features. Multiple classifiers are selected so that various forms of the same character can be considered as the distinct classifiers generate different results for the similar features. So, the character recognition system's output is the label that identifies each character or symbolic representation. Fig. 2 demonstrates the flow chart of the proposed character recognition system. Initially, the database of handwritten characters is prepared with variations in the font style and font size. After creating the database, isolated the characters in an automated fashion. These characters are preprocessed to remove the noise. Features are extracted from the preprocessed data and these characters will be input for the various ML classifiers.

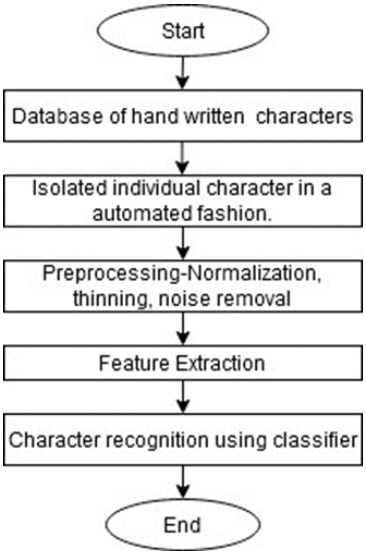
Start
Database of hand written characters
Isolated individual character in a automated fashion.
Preprocessing-Normalization, thinning, noise removal
Feature Extraction
Character recognition using classifier
End

Fig. 2 Flow chart character recognition system

The dataset of handwritten characters is taken which is preprocessed to remove the redundant information from the images. To make the characters scale-invariant aspect ratio adaptive normalization is used. The standard plane of fixed size 32 × 32 or a normalized image plane with side length S is used in this technique. The original image's aspect ratio is calculated using equation (1)

$$A1 = \min(w_1,h_1)/\max(w_1,h_1) \qquad (1)$$

The fixed aspect ratio mapping is used to compute the aspect ratio of normalized image. On the standard plane with the dimension $S = \max(w_2,h_2)$, the normalized image centered. The $\min(w_2,h_2)$ is calculated using equation (2)

$$A2 = \min(w_2,h_2)/\max(w_2,h_2) \qquad (2)$$

Normalization, thinning and noise removal is implemented in the preprocessing step to eliminate extra pixel information. After preprocessing, the features of the characters are extracted and then various classifiers like Decision Tree classifier, Nearest Centroid classifier, Kneighbors classifier, ExtraTrees classifiers and Random Forest classifier are used to recognize the characters. To obtain the scores of these algorithms and to compare their performance, the grid search is performed. Using grid search optimum parameters can be obtained.

IV. Results and Discussions

This section discusses the results deduced from selected classifiers to recognize the characters in the character recognition system. Further, the performance of various classifiers is analyzed. For the entire selected Machine Learning (ML) classifiers such as Decision Tree classifier, Nearest Centroid classifier, K Nearest Neighbors classifier, Extra Trees classifiers and Random Forest classifier accuracy is compared. Table 1 shows the tabulated results of the accuracy comparison of various ML classifiers. Among all the classifiers Extra Trees classifiers and Random Forest classifier showed the best with the accuracy of 78.19% and 76.82% respectively.

Table I – Accuracy of various classifiers

Algorithm	Accuracy (in percent)
DTC	63.51
NC	68.02
KNC	75.14
ETC	78.19
RFC	76.82

Fig. 3 is a graphical representation of the accuracy versus various ML classifiers used in the system. The Extra Trees Classifier and Random Forest proved the best among all the algorithms used.

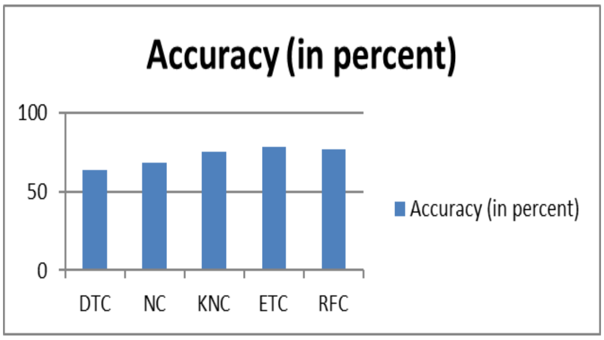

Fig. 3 Accuracy vs Classifiers

Fig. 4 shows the accuracy for K Nearest Neighbors Classification algorithm. The accuracy was less when the number of neighbors (K) considered 20.

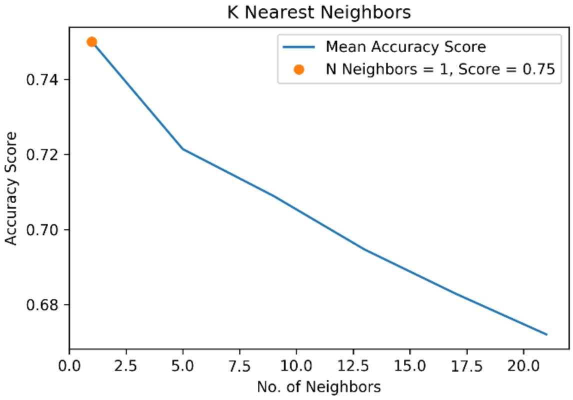

Fig. 4 Nearest Neighbors Classification Algorithm Accuracy

Fig. 5 demonstrates the accuracy of the Extremely Randomized Decision Forest Classification algorithm. As the number of trees increased, the accuracy also increased. The highest accuracy is recorded for the total number of trees equal to 280. This is the threshold value for the total number of trees used to compose the Random forest. After this threshold value, though the number trees increased, there would be no significant change in the gain value.

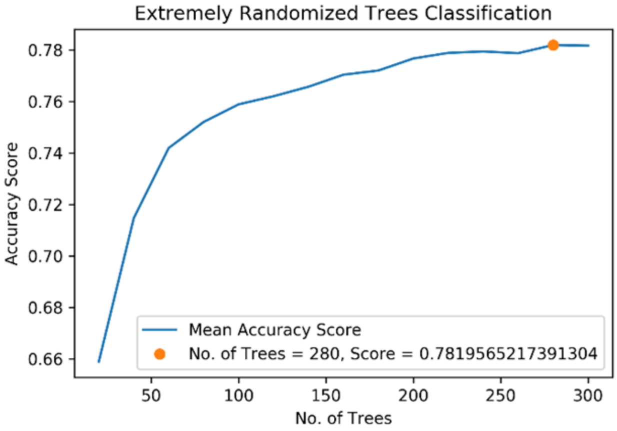

Fig. 5 Extremely Randomized Decision Forest Classification Algorithm Accuracy

Fig. 6 presents the Random Forest Classification Algorithm Accuracy. The accuracy is calculated starting from the number of trees equal to 10 till 300. The highest accuracy 76.82 is obtained at 260 the number of trees chosen.

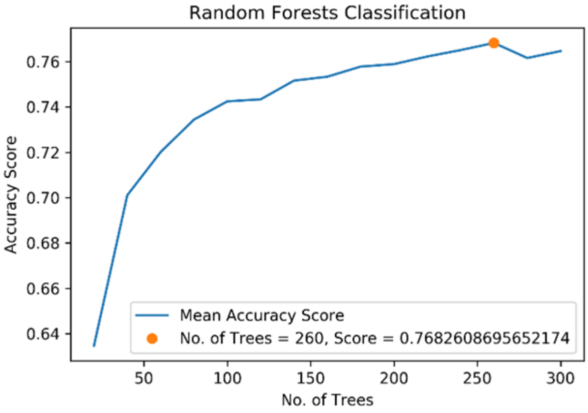

Fig. 6 Random Forest Classification Algorithm Accuracy

V. CONCLUSION

This paper highlighted the drawbacks of manual processing of the handwritten data. Then stride through the proposed model which implemented a character recognition system for Devanagari script using various machine learning classifiers like Decision Tree classifier, Nearest Centroid classifier, K Nearest Neighbors classifier, Extra Trees classifiers and Random Forest classifier. The results obtained for all classifiers clearly envisaged the Extra Trees classifiers and Random Forest Regression as the best among the used classifiers. The statistical measure technique used for comparison was the accuracy of classifiers. The future work will implement the hybrid approach to increase the accuracy of the model.

VI. ACKNOWLEDGEMENT

The author is thankful to all authors who have participated to conduct the experiment and research in the proposed research work. We are thankful to the CMR Institute of Technology for providing all resources and computing facilities to conduct this research.

REFERENCES

[1] Mohmed Cheriet, Nawwaf Kharma, and Cheng-Lin Liu, “Character Recognition Systems- A Guide for Students and Practioners,” 3rd ed. vol. 2, Wiley-interscience, 2007, pp.302-303.

[2] Xiaolin Li, and Dit-yan Yeung, “On-line handwritten alphanumeric character recognition using dominant points in strokes,” Journal of Pattern Recognition, vol. 30, No. 1, pp. 31-44, 2011.

[3] Gupta D, and Bag S. “An efficient character segmentation approach for handwritten hindi text,” In 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN) 2018 Feb 22 (pp. 730-734). IEEE.

[4] Deokate, S. T., and N. J. Uke, “Hybrid methods for Segmenting and Identifying the Marathi Text,” 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). IEEE, 2019.

[5] Veena Bansal, and R.M.K. Sinha, “Integrating knowledge sources in Devnagari Text recognition system,” IEEE Transactions on systems, Man, and cybernetics-Part-A: systems and humans. Vol.30 no 4, pp 500-505, July 2000.

[6] Roy A, Bhowmik TK, Parui SK, and Roy U, “A novel approach to skew detection and character segmentation for handwritten Bangla words,” Digital Image Computing: Techniques and Applications”, (DICTA'05), IEEE. 2005 Dec 6, (pp. 30-30).

[7] Seethalakshmi R, Sreeranjani TR, Balachandar T, Singh A, Singh M, Ratan R, and Kumar S. “Optical character recognition for printed Tamil text using Unicode ,” Journal of Zhejiang University-SCIENCE A. 2005 Nov;6(11):1297-305.

[8] X. Du, W. Pan, and Tien D. Bui, “Text line segmentation in handwritten documents using Mumford–Shah model,” Pattern Recognition. Vol. 42, pp. 3136 – 3145, 2009.

[9] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, “Text line and word segmentation of handwritten documents,” Pattern recognition Vol. 42, pp. 3169-3183, 2009.

[10] Saha S, Basu S, Nasipuri M, and Basu DK. “A Hough transform based technique for text segmentation,” arXiv preprint arXiv:1002.4048. 2010 Feb 22.

[11] Singh B, Gupta N, Tyagi R, Mittal A, and Ghosh D, “Parallel implementation of Devanagari text line and word segmentation approach on GPU. Jayadevan, Offline Recognition of Devanagari Script: A Survey,” IEEE Transactions International Journal of Computer Applications. 2011 Jun; 24(9):7-14.

[12] Deepti Khandja, Neeta Nain, and Subhash Panwara, “Hybrid feature extraction Algorithm for Devanagari script,” ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) , vol. 15, article no.2, pp. 605-615, November 2015.

[13] R. Jayadevan, ”Offline recognition of Devanagari script: a survey,” IEEE Trnsactions on Systems, Man, and CybernaticsPart C: Applications and Reviews, vol. 41, no. 6, pp. 782-796, November 2011.

[14] Gyanendra K. Verma, Shitala Prasad, Piyush Kumar and C. Singh, “Handwritten Hindi character recognition using curvelet transform,” International Conference on Information Systems for Indian Languages, Springer, pp. 224-227, 2011.

[15] L.Malik, and Dr. P.S. Deshpande , “Recognition of handwritten Devnagari script,” International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI) , vol. 24, no. 5, pp.1-14, 2010.

[16] A. Desai and L. Malik, “A modified approach to thinning of Devanagri characters,” International Conference on Electronics Computer Technology (ICECT), vol. 1, pp. 420-423, April 2011.

[17] R. W. Zhou, C. Quek, and G. S. Ng, “A novel single-pass thinning algorithm and an effective set of performance criteria,” Journal of Pattern Recognition Lett.,vol. 16, no. 12, pp. 1267-1275, Dec. 1995.

[18] B. Thakral and M. Kumar, “Devanagari handwritten text segmentation for overlapping and conjunct characters- a proficient technique,” International Conference on Reliability, Infocom Technologies and Optimization,pp. 1-4, Oct 2014.

[19] Anupama Ray, Ankit Chandawala and Santanu Chaudhury, “Character recognition using conditional andom field based recognition engine,” International Conference on Document Analysis and Recognition (ICDAR) , pp. 18-22, 2013.

[20] M. Jangid, R. Dhir, R. Rani, and K. Singh, “SVM classifier for recognition of handwritten Devanagari numeral,” International Conference on Image Information Processing (ICIIP),IEEE, pp. 15, Nov 2011.

[21] Avadesh M, and Goyal N, “Optical character recognition for sanskrit using convolution neural networks,” 13th IAPR International Workshop on Document Analysis Systems (DAS),IEEE, pp. 447-452, Apr 2018.

[22] Karpe RP, and Vernekar N, “Konkani script to speech conversion by concatenation of recognized hand written Konkani text using neural network,” International Conference on Computer Communication and Informatics (ICCCI),IEEE, pp. 1-6, Jan 2019.

[23] Sonth S, and Kallimani JS, “OCR based facilitator for the visually challenged,“ International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT) , IEEE, pp. 1-7, Dec 2017.

[24] Todase JA, and Shelke S, “Script translation system for Devnagari to English,” International Conference on Information, Communication, Engineering and Technology (ICICET) , IEEE, pp. 1-4, Aug 2018.

[25] Farkya S, Surampudi G, and Kothari A, “ Hindi speech synthesis by concatenation of recognized hand written devnagri script using support vector machines classifier,” International Conference on Communications and Signal Processing (ICCSP), IEEE, pp. 0893-0898, April 2015.

[26] Jacob IJ. “Performance evaluation of caps-net based multitask learning architecture for text classification,” Journal of Artificial Intelligence. 2020;2(01):1-0.

[27] Vijayakumar T, Vinothkanna MR, “Capsule Network on Font Style Classification,” Journal of Artificial Intelligence. 2020 May;2(02):64-76.