WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Recognition of Similar Handwritten Hindi Characters

[1]Sai Chaitanya. K, [2]Sai Charan P, [3]Sai Charan. R, [4]Sai Harshitha. I, [5]Sai Kamal. M, [6]Sai Kiran Kumar Reddy. C

[1](2111cs020425), [2](2111cs020426), [3](2111cs020427), [4](2111cs020428), [5](2111cs020429), [6](2111cs020430)
Artificial Intelligence and Machine Learning Department, Mallareddy University, Hyderabad, Telangana, India

ABSTRACT:

This paper explores the application of Machine Learning in the domain of Handwritten Hindi Characters Recognition. With the growing need for efficient and accurate character recognition systems, especially in languages with complex scripts like Hindi, the study focuses on addressing the challenge of distinguishing similar characters in handwritten forms. Experimental results demonstrate the effectiveness of the method in achieving high accuracy and robust performance in recognizing handwritten Hindi characters that exhibit visual similarities. Handwritten character recognition plays an important role and possible applications in assisting technology for blind and visually impaired users, human–robot interaction, automatic data entry for business documents, etc.

Introduction:

Optical Character Recognition (OCR) is a field of research in pattern recognition, artificial intelligence and machine vision. OCR is a mechanism to convert machine printed or handwritten document file into editable text format. In handwritten Character Recognition, there are lots of problems as compare to machine printed document because of the different peoples have different writing styles, the size of pen-tip and some people have skewness in their writing. All this challenges make the researches to solve the problems.

Devnagari Script is an oldest one that is used to write many languages such as Hindi, Nepali, Marathi, Sindhi and Sanskrit where Hindi is the third most popular language in the world and it is the national language of the India [1]. 300 million people use the Devnagari Script for documentation in central and northern parts of India [2].

First research report on handwritten Devanagari Characters was published in 1977[3] after which many researchers have done the work on the Devanagari Script with different feature extraction algorithms and different classifiers. G S Lehal and Nivedan Bhatt [4] have proposed a contour extraction technique and obtained 89% accuracy. Reena Bajaj et al. [5] have employed three different kinds of feature namely, density features, moment features and descriptive features for classification of Devanagari Numerals and obtained 89.68% accuracy.

Limitations:

- **Ambiguity in Characters :** Some Hindi characters might look similar, especially in handwritten text, leading to ambiguity. For instance, characters like र (ra) and च (cha) might be challenging to distinguish if written hastily
- **Noise And Distortions:**Noisy or distorted input, due to factors such as smudging, uneven ink flow, or creases in the paper, can negatively impact the recognition accuracy.
- **Lack Of Standardisation in Handwriting:** Unlike printed text, handwriting lacks a standardized format. This lack of uniformity makes it challenging to create a one-size-fits-all recognition system.

Proposed Work

- In this work, we propose a technique to recognize handwritten Hindi characters using deep learning approaches like Convolutional Neural Network (CNN).This is a Character Recognition System which we developed for Devanagari Script.
- We try to create a hindi character identifier, i.e A machine has to identify the hindi characters and can able to recognose them and provide accurate output which helps in many ways

DATABASE:

The database is provided by the ISI (Indian Statistical Institute, Kolkata) [14]. Initially Devanagari script was developed to write Sanskrit but was later adapted to write many other languages such as Hindi Marathi and Nepali. The printed Devanagari Numerals are shown in figure 1 and it is seen that there are variations in the shapes of numerals 5, 8 and 9 in their printed forms. In figure 2, there are shown the samples of the Handwritten Devanagari Numerals database. The distributions of training data and testing data are shown in table 1.

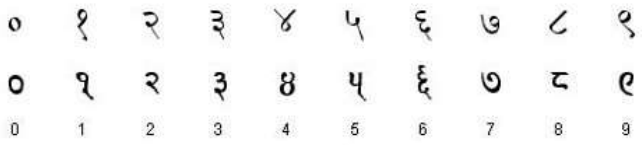

Figure 1: Devanagari Numerals

Table 1
Brief details of classification techniques used for Devanagari manuscript recognition.

Data set	Classification Technique	Feature Extraction Technique	Accuracy	Citation
3500 characters	CNN	CNN	93%	Narang et al. (2021b)
30,305 characters	DCNN	DCNN	96.02%	Jangid and Srivastava (2018)
6152 characters	NA	Statistical feature extraction techniques	88.95%	Narang et al. (2019a)
9956 characters	NA	SIFT and Gabor Features	91.29%	Narang et al. (2021c)
24500 characters	K-NN and Linear SVM	Zoning, Diagonal, transition and peak extent	91.53%	Narang et al. (2020)
5484 characters	RBF-SVM Classifier	Ada Boost and Bagging ensemble	91.70%	Narang et al. (2021a)

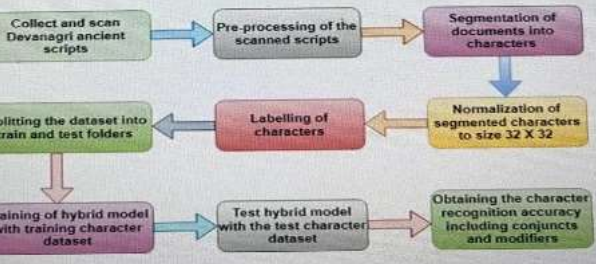

Fig. 4. Block diagram of the proposed model.

Data Set Descriptions :

To perform handwritten Hindi character recognition using Support Vector Machines (SVM), you'll need a labeled dataset that includes images of handwritten Hindi characters along with corresponding labels indicating the identity of each character. Here's a description of the required dataset:

.Dataset Format:

- Images: The dataset should consist of images of handwritten Hindi characters. Each image should be in a standardized format, such as JPEG or PNG.
- Labels: Corresponding labels for each image indicating the identity of the handwritten Hindi character.

Character Set:

- Define the set of Hindi characters you want to recognize. This may include basic characters, consonants, vowels, and other symbols commonly used in the Hindi script.

Image Size and Quality:

- Ensure that all images are of the same size and quality. Consistency in image dimensions helps in preprocessing and feeding the data into the SVM model.

Training and Testing Sets:

- Split the dataset into training and testing sets. A common split is to use a majority of the data for training (e.g., 70-80%) and the rest for testing to evaluate the model's performance

Implementation Details:

- Choose a suitable SVM implementation library, such as Scikit-learn in Python. Train the SVM model using the training dataset and evaluate its performance on the testing dataset.

Dataset Source and Licensing:

- We have collected data from Devanagari Handwritten Manuscript Dataset.
- Devanagari is a collection of 11 vowels and 33 consonants and 3 conjuncts. Vowels are independent characters or they can also be written using diacritical marks which are known as modifiers. The characters that are formed by combining modifiers are known as conjuncts.
- Two or more consonants combine to form a compound conjunct. The horizontal line on the uppermost part of the compound conjunct is known as Shirorekha. Devanagari elements and sample words are shown in Fig.

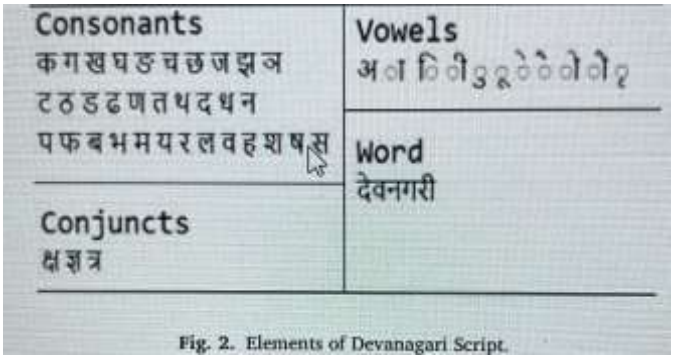

Fig. 2. Elements of Devanagari Script.

Fig. 3. Sample image of Devanagari Manuscript.

- A Devanagari manuscript has multiple words written either handwritten or typed. A sample image is shown in Fig. 3

Research problem and contribution :

- Regional languages have high difficulty levels due to the presence of basic characters, conjunct, modifiers, etc. So, regional language based manuscripts have additional complexity levels due to divided Shirorekha, transitive text, bilingual text, different starting points, Orientation of characters, etc.
- The research shows that deep networks perform better as compared to shallow ones as they are capable of handling high-dimensional and complicated data. Although the ability of deep learning models has not been fully utilized for the field of handwriting recognition . Some variants of deep learning include Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN) is used to generate the handwriting using the sample images. Further generated data was used to classify the character.

Data Preprocessing Techniques:

Preprocess the images to enhance the SVM model's performance. Common preprocessing steps include resizing, normalization, and converting images to grayscale.

Image annotation and Segmentation:

Initially, there is a requirement of segmenting the image into characters. Manuscript documents are annotated and segmented into characters using the Vott tool. Vott is an open source for annotating images that helps us even to write scripts for complex code segmentation.

This tool takes images as input and generates one JSON file per image. The segmented character images can, later on, be used to form classes for the character recognition process.

Caps Net Model:

Caps Net is a type of neural network that comprises a collection of neurons whose activity vector indicates the instantiating parameters of a particular entity, such as an object or an object portion

Model Selection-Model Development :

Support Vector Machine:

The accuracy of the Support Vector Machine Algorithm is 7.229%

Optical Character Recognition:

Optical Character Recognition is a technique that is used to convert scanned images like handwritten text into an editable format.

The basic steps in the OCR is Manuscript, Preprocessing, Segmentation, Feature Extraction, Classification and

Prediction And Analysis

Convolution Neural Networks:

- Variant of deep learning
- Learning technique is used to generate the handwriting using the sample images

Recursive Neural Networks:

- Machine Learning technique is used to generate the handwriting using the sample images
- **Tensor Flow:** TensorFlow is an end-to-end open source platform for machine learning. TensorFlow API to develop and train machine learning models.
- **Keras:**Keras is a high-level, deep learning API developed by Google for implementing neural networks.
- It also supports multiple backend neural network computation.

Model Training and Testing:

One half for model training and also the other part for model analysis or testing. During this study, the info set is separated into two part the first half is termed coaching knowledge and also the second called take a look at data, training data makes up for eighty percent of the whole data used, and the rest for test data. all of those models are trained with the training data part and so evaluated with the test data.

CONCLUSION

In this paper a feature extraction algorithm has been proposed to recognize handwritten Devanagari Numerals. The results for recognition of test data and 5 fold cross validation of whole dataset are observed. 32*32 normalized size of samples producing 144 features is efficient producing 98.62% test recognition rate and 98.94% cross validation accuracy. For improvement multiple classifiers can be combined with additive recognition rate with diverse features.

REFERENCES

[1] Journals and publications dedicated to image processing, pattern recognition, and machine learning. Journals like the International Journal of Computer Applications or Pattern Recognition Letters.

[2] Books that cover the fundamentals of Support Vector Machines and their applications in image processing.

[3] G S Lehal, Nivedan Bhatt, “A Recognition System for Devnagri and English Handwritten Numerals”, Proc. Of ICMI, 2000.

[4] Reena Bajaj, Lipika Day, Santanu Chaudhari, “Devanagari Numeral Recognition by Combining Decision of Multiple Connectionist Classifiers”, Sadhana, Vol.27, Part-I, 59-72, 2002.

[5] R.J.Ramteke, S.C.Mehrotra, “Recognition Handwritten Devanagari Numerals”, International journal of Computer processing of Oriental languages, 2008.

[6] U. Bhattacharya, S. K. Parui, B. Shaw, K. Bhattacharya, “Neural Combination of ANN and HMM for Handwritten Devnagari Numeral Recognition”.

[7] U.Pal, R.K.Roy and F. Kimura,”Indian muilti-script full pin-code string recognition for postal automation” in Proc. 10 th conf. Document Analysis Recognition, 2009, pp 456-460

[8] Some information from: (https://scholar.google.com/) and https://github.com/ and conferences like ICDAR (https://www.icdar2021.org/) and CVPR (https://cvpr2023.thecvf.com/).