Hand-written Hindi Word Recognition - A Comprehensive Survey

Pooja Patel[1], Bhavesh Tanawala[2], Pranay Patel[3]

[1]Student (Mtech), Department of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Anand, Gujarat, India
[2]Professor, Department of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Anand, Gujarat, India
[3]Professor, Department of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Anand, Gujarat, India

---

Abstract- *Word recognition is a technique that allows different types of scanned documents to be transformed into searchable and editable data. It is divided into two main categories: printed word recognition and handwritten word recognition. Many academics have already been working on word recognition for a decade. This work gives a detailed review of known strategies for hand-written word recognition in Hindi scripts. This work seeks to give scholars in the field of hand-written Hindi word recognition some insight.*

Key Words: Hand-written Hindi Word recognition, Feature extraction, Classification

1. INTRODUCTION

Hand-written word recognition is a difficult and active topic for researchers in the subject of pattern recognition. Several optical character recognition (OCR) systems have been created for this purpose and are being utilized in numerous commercial applications, including mail reading help for the blind, automatic number plate identification, form processing, bank cheque processing, and postal address recognition [1].

Because of the differences in the forms of letters, it is difficult for a machine to identify/recognize a handwritten character. These differences are caused by the writer's diverse writing style, the actuation device, the pen width, the ink color, and a variety of other characters, (2) their structure significant barrier in designing the system for hand-written Hindi words recognition. And shape, and (3) similar shaped characters, hand-written factors. Furthermore, because of (1) a large word set with more curves, loops, and other intricacies in Hindi characters are difficult to recognize.

As a result, researchers have such the remaining part of the paper is written out as follows: Hindi script along with its properties is discussed in section 2. The working model is written in section 3. Observation is discussed in section 4. The parametric evaluation of existing techniques for Hand-written character recognition is presented in Section 5. We conclude the paper in section 6.

2. HINDI SCRIPT AND ITS PROPERTY

There are 12 vowels, 36 consonants, and 10 number characters in the Hindi script [3]-[6]. A vowel can be written as a single character or as a combination of characters formed by combining Mantras. Barakhadi characters are made up of a mix of vowels and consonants. Shiro Rekha, or headline, is a horizontal line that appears at the top of every character in Hindi [4]-[6].

From left to right, Hindi is written. It does not have upper- and lower-case letters like the English language. Furthermore, the Hindi character set has more symbols than the English character set. Curves, holes, and strokes make up the majority of characters in Hindi writing, with the help of the Shiro Rekha. Table 1-4 shows the Hindi character set.

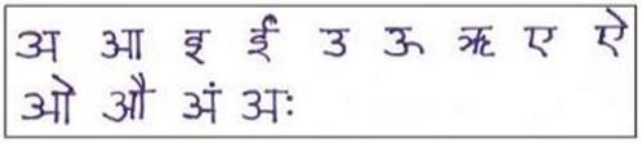

TABLE -1: VOWEL CHARACTER [2]

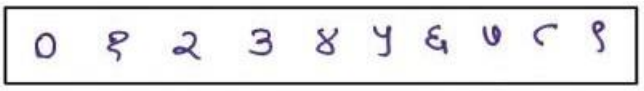

TABLE -2: NUMBER CHARACTER [2]

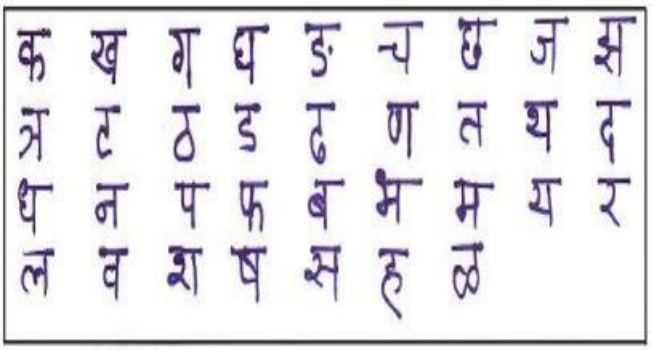

TABLE -3: CONSONANT CHARACTER [2]

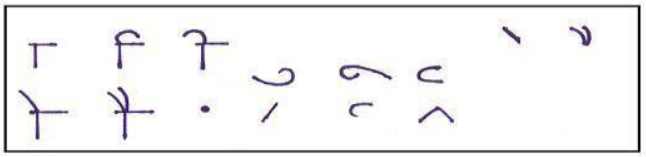

TABLE -4: MODIFIERS [2]

3. WORKING MODEL

The stages that describe how the recognition of Hand-Written Hindi Word works are as follows: (See Figure 1)

A. Data Acquisition

Handwritten Hindi texts are scanned and appropriate digital images are prepared during data capture. The images are then sent on to the next stage of pre-processing (Fig. 1).

B. Pre-processing

The initial phase in the character recognition process is pre-processing. It takes real-time photos and transforms them into a unique format before removing noise and undesired background. Noise reduction [12],[17], binarization [12],[18], normalization, skew correction, and other pre-processing techniques are used. The subsequent phases, feature extraction, and classification are dependent on the pre-processing step, as mentioned in.

C. Segmentation

input images are normalized during the pre-processing stage. Words and characters are retrieved from images during the segmentation process. The purpose of segmentation is to divide an image into a group of disjoint sections that are distinct and meaningful in terms of several attributes.

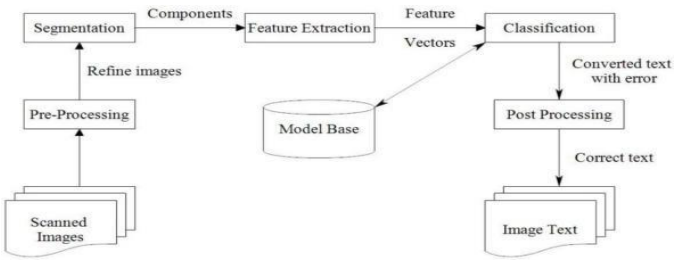

Fig -1: Working of a system[25]

D. Feature Extraction

Feature extraction is a crucial stage in OCR before categorization (Fig. 1). The extracted characteristics have a significant influence on the accuracy of Text recognition. Specific features (characteristics) are retrieved and saved in a feature vector for all input images in this stage. These characteristics are separated into two categories: statistical characteristics and structural characteristics [8].

Local or worldwide statistical characteristics are available. Global features are taken from the whole character image, whereas local features are extracted from the image's immediate vicinity. Some of the most common statistical characteristics are moments, zoning, projection, histogram, n-tuples, crossing, and distances. Noise and distortions also have little effect on global or local characteristics [8].

Topological characteristics define both global and local qualities of pictures, and structural features are based on them. Object components, structures, and attributes are described using topological features. Loops, end-points, extreme points, intersections, and other topological characteristics are common examples [8].

E. Classification

The categorization phase, as illustrated in Fig. 1, is the next stage in the recognition system's operation, in which the first character is recognized. Following that, the identified word is assigned to a preset class.

F. Post Processing

The post-processing stage is the system's final step. This step's main purpose is to output the relevant detected character as structured text [9].

4. OBSERVATION

The authors of [11] used neural networks to recognize handwritten Devanagri special letters and words. Aside from an input layer and an output layer, the neural classifier has two hidden layers. The eight-neighbor nearby approach is used to retrieve information about a handwritten character's boundaries. For special characters in the Devanagari script, the proposed technique has produced an accuracy of up to 90%.

The authors of [12] described an offline SVM approach for handwritten Hindi text recognition. For each feature, multiple algorithms based on the form of the character we used to extract shape-based characteristics. During the feature selection step, a total of 59 features are chosen. The recognition rate is 89.6 %.

The authors of [13] used Artificial Neural Networks to recognize handwritten Devanagari signatures in real-time (ANN). Different aspects of the signature are retrieved and utilized to train the Neural Network, such as height, slant, and length. The authors gathered 500 valid signatures in all. The suggested Devanagari handwritten signature recognition system had a 96.12 % accuracy rate.

The authors of [14] create and consider a new Devanagari Hand-written Character Dataset for the Devanagari script. This collection contains 92,000 pictures of 46 distinct classes of Devanagari script characters. They perform pre-processing tasks such as image scaling, padding, and grayscale conversion. For character recognition, they use the CNN classifier.

The researchers of [17] take into account a database made up of hand-written Hindi characters. It includes of total 4,428 samples and 108 samples for each character with an intent to assure diverse orientations and sizes. They employ the Histogram Oriented Gradient (HOG) and the Profile Projection Histogram. (PP) is a program that extracts features. They employ a variety of classifiers-Ensemble Subspace Discriminant, K-NN, weighted K-NN, bagged trees, quadratic SVM.

The authors of [12] utilize region-based k-means clustering for character feature extraction. The picture is binarized by them. Character separation is done at the pre-processing stage. They make use of a 430-sample Hindi characters database. They also use SVM and Euclidean distance to classify the data.

The creator of [18] utilizes a database that has 10 example images for each of the 62 Handwritten characters. Each image is 80x40 pixels in size. They employ a Histogram Oriented Gradient (HOG) to extract features. For character categorization, they employ an Artificial Neural Network (ANN).

The authors of [19] propose a new method for recognizing Hindi characters based on the digital curvelet transform and the K-Nearest Neighbour (KNN) classifier. Curvelet features are retrieved by calculating thick and thin images using the curvelet transform after the input images have been segmented. They take into account the database, which contains 200 images of the character set.

The authors present a paradigm for handwritten number recognition in Indic scripts in [21]. For error reduction, the model uses CNN with backpropagation and dropout for data overfitting. They look at the Bangla hand-written numerical dataset, Urdu hand-written numerical dataset, and Hindi hand-written numerical dataset, which are all in three languages.

The writers evaluate a database of Devanagari's old manuscripts from various libraries and museums in [24][4]. The characters are normalized to 64x64 pixels. Using the nearest-neighbor interpolation approach, you may create a unique character. The authors apply a Discrete Cosine Transform. In [24] (DCT) zigzag is used to extract features. Classifiers such as decision trees [22], Support Vector Machines (SVM) [22], and Naive Bayes [22] are used to recognize the character in manuscripts. To increase accuracy, the basic classifier is combined with Ada boost and bagging ensemble algorithms. Adaptive boosting using Radial Basis Function (RBF) kernel SVM achieves the highest accuracy. The authors analyze the quality of the ensemble process using a variety of performance evaluation metrics. The authors of[4] extract feature using a zigzag Discrete Cosine Transform (DCT) and Histogram of Oriented Gradients (HOG) [7],[23]. For character identification from ancient Devanagari manuscripts, decision trees, Support Vector Machines, and Naive Bayes classifiers are utilized.

5. PARAMETRIC EVALUATION

The parametric assessment of available approaches for handwritten character recognition is Pre-processing approaches, feature extraction techniques, classification techniques, languages examined, performance measurements, number of samples utilized, number of classes, and types of characters studied are all employed for this purpose. Here the paper suggested the Hand-written Hindi word recognition technique from author to author, it is vary.

6. CONCLUSIONS

We give a complete assessment of character and numeric identification work for hand-written Hindi scripts in this paper. The study discusses various methods for recognizing hand-written words in Hindi script from images. It also includes a list of the most frequent feature extraction and pre-processing approaches. This paper also includes a parametric assessment of these current algorithms, which will be valuable to researchers working on hand-written Hindi character identification.

7. REFERENCES

[1] Malanker, A. and Patel, P., 2014. Handwritten Devanagari script recognition: a survey. IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) e- ISSN, pp.2278-1676.Malanker, A. and Patel, P., 2014.

[2] Patil, P.M. and Ansari, S., 2013. A research survey of devanagari handwritten word recognition. Int. J. Eng. Res. Technol, 2(10), pp.1010-1015.

[3] Gaur, A., and Yadav, S., 2015, January. Handwritten Hindi character recognition using k-means clustering and SVM. In 2015 4th international symposium on emerging trends and technologies in libraries and information services (pp. 65-70). IEEE.

[4] Narang, S.R., Jindal, M.K. and Sharma, P., 2018, December. Devanagari ancient character recognition using HOG and DCT features. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 215-220). IEEE.

[5] Acharya, S., Pant, A.K. and Gyawali, P.K., 2015, December. Deep learning-based large-scale handwritten Devanagari character recognition. In 2015 9th International conference on software, knowledge, information management and applications (SKIMA) (pp. 1-6). IEEE.

[6] Verma, G.K., Prasad, S. and Kumar, P., 2011, March. Handwritten Hindi character recognition using curvelet transform. In International Conference on Information Systems for Indian Languages (pp. 224-227). Springer, Berlin, Heidelberg.

[7] Yadav, M. and Purwar, R., 2017, January. Hindi handwritten character recognition using multiple classifiers. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence (pp. 149-154). IEEE.

[8] Yadav, M., Purwar, R.K. and Mittal, M., 2018. Handwritten Hindi character recognition: a review. IET Image Processing, 12(11), pp.1919-1933.

[9] Jain, A.K., Duin, R.P.W. and Mao, J., 2000. Statistical pattern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence, 22(1), pp.4-37.

[10] Singh, B., Mittal, A., Ansari, M.A. and Ghosh, D., 2011. Handwritten Devanagari word recognition: a curvelet transform based approach. International Journal on Computer Science and Engineering, 3(4), pp.1658-1665.

[11] Dedhe, V. and Patil, S., 2013. Handwritten Devanagari special characters and words recognition using neural network. Int. J. Eng. Sci. Res. Technol, 2(9), pp.2521-2526.

[12] Garg, N.K., Kaur, L. and Jindal, M., 2013. Recognition of offline handwritten Hindi text using SVM. International Journal of Image Processing (IJIP), 7(4), pp.395-401.

[13] Dewangan, S.K., 2013. Real-Time Recognition of Handwritten Devnagari Signatures without Segmentation Using Artificial Neural Network. International Journal of Image, Graphics and Signal Processing, 5(4), p.30.

[14] Acharya, S., Pant, A.K. and Gyawali, P.K., 2015, December. Deep learning-based large scale handwritten Devanagari character recognition. In 2015 9th International conference on software, knowledge, information management and applications (SKIMA) (pp. 1-6). IEEE.

[15] Narang, S.R., Jindal, M.K. and Kumar, M., 2019. Devanagari ancient character recognition using DCT features with adaptive boosting and bootstrap aggregating. Soft Computing, 23(24), pp.13603-13614.

[16] Narang, S.R., Jindal, M.K. and Sharma, P., 2018, December. Devanagari ancient character recognition using HOG and DCT features. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 215-220). IEEE.

[17] Yadav, M. and Purwar, R., 2017, January. Hindi handwritten character recognition using multiple classifiers. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence (pp. 149-154). IEEE.

[18] Singh, N., 2018, February. An efficient approach for handwritten devanagari character recognition based on an artificial neural network. In 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 894-897). IEEE.

[19] Verma, G.K., Prasad, S. and Kumar, P., 2011, March. Handwritten Hindi character recognition using curvelet transform. In International Conference on Information Systems for Indian Languages (pp. 224- 227). Springer, Berlin, Heidelberg.

[20] Tushar, A.K., Ashiquzzaman, A., Afrin, A. and Islam, M., 2018. A novel transfer learning approach upon Hindi, Arabic, and Bangla numerals using convolutional neural networks. In Computational Vision and Bio-Inspired Computing (pp. 972-981). Springer, Cham.

[21] Tushar, A.K., Ashiquzzaman, A., Afrin, A. and Islam, M., 2018. A novel transfer learning approach upon Hindi, Arabic, and Bangla numerals using convolutional neural networks. In Computational Vision and Bio-Inspired Computing (pp. 972-981). Springer, Cham.

[22] Mahesh, B., 2020. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, pp.381-386.

[23] Dalal, N. and Triggs, B., 2005, June. Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). Ieee.

[24] Narang, S.R., Jindal, M.K. and Kumar, M.,2019. Devanagari ancient character recognition using DCT features with adaptive boosting and bootstrap aggregating. Soft Computing, 23(24), pp.13603-13614.

[25] Singh, A.K., Kadhiwala, B. and Patel, R., 2021, October. Hand-written Hindi Character Recognition-A Comprehensive Review. In 2021 2nd Global Conference for Advancement in Technology (GCAT) (pp. 1- 5). IEEE.