2021 2nd Global Conference for Advancement in Technology (GCAT)
Bangalore, India. Oct 1-3, 2021

Hand-written Hindi Character Recognition - A Comprehensive Review

Awadh Kishor Singh
PG Student, Computer Engineering Department,
Sarvajanik College of Engineering and Technology
Surat, India
awadhks97@gmail.com

Bintu Kadhiwala
Assistant Professor, Computer Engineering Department,
Sarvajanik College of Engineering and Technology
Surat, India
bintukadhiwala@gmail.com

Rakesh Patel
Assistant Professor, Computer Engineering Department,
Sarvajanik College of Engineering and Technology
Surat, India
rakeshpatel.ce@gmail.com

2021 2nd Global Conference for Advancement in Technology (GCAT) | 978-1-6654-1836-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/GCAT52182.2021.9587554

***Abstract*—Character recognition is a technology that facilitates the conversion of different types of scanned documents into searchable and editable data. Since the decade of years, many researchers work on character recognition. It can be classified into hand-written character recognition and printed character recognition. Hand-written character recognition is considered to be a demanding research area in the field of pattern recognition. In this paper, we present a comprehensive survey on existing techniques for hand-written character recognition of Hindi scripts with the help of various parameters such as techniques utilized for pre-processing, feature extraction, classification, etc. This paper aims to provide an insight to researchers working in the domain of hand-written Hindi character recognition.**

***Index Terms*—Hand-written character recognition, OCR, Feature extraction, Classification**

I. Introduction

In the pattern recognition domain, Hand-written character recognition is an active and challenging research area. For the purpose, several Optical Character Recognition (OCR) systems are developed and used in various commercial applications such as mail reading assist for the blind, automatic number plate recognition, form processing, bank cheque processing, and postal address recognition [1], [2].

It is difficult for a machine to identify/recognize a Hand-written character because of variations in the shapes of characters. These variations are due to writer's wide-ranging writing style, the actuation device, the pen width, the ink colour, and many other factors. In addition, hand-written Hindi characters are complex to recognize because of, (1) a large set of characters having more loops, curves, and further details of the characters, (2) their structure and shape, and (3) similar shaped characters. Hence, developing a system for the hand-written Hindi character recognition poses a major challenge to the researchers.

The rest of the paper is organized as follows: An Optical Character Recognition process is discussed in section II. Hindi script along with its properties is discussed in section III. Related work is discussed in section IV. The parametric evaluation of existing techniques for Hand-written character recognition is presented in Section V. We conclude the paper in section VI.

II. Optical Character Recognition

The scanned images of either hand-written or machine-printed text are converted into a computer format text. This coversion process is termed as Optical Character Recognition (OCR) [3]. Typically, the OCR system examines the scanned image of the document, identifies all characters and/or numbers, and generates text in a machine readable form that can be used by the computer system. In other words, the OCR system takes a scanned image of the document as an input and generates a text document as an output.

As discussed in [2], based on data acquisition process, the OCR is classified into (1) off-line character recognition, and (2) online character recognition. Online character recognition converts input text into digital text as soon as it is entered through the devices like mobile, pad, or any digitizer. Off-line character recognition converts hand-written or machine-printed text images into the digital text. The off-line character recognition is further classified into (1) machine-printed character recognition, and (2) hand-written character recognition [2]. The working of the OCR is described by the following steps (Fig. 1):

A. Data Acquisition

In data acquisition, hand-written character documents are scanned and corresponding digital images are generated. Then, these images are passed to the next pre-processing phase (Fig. 1).

B. Pre-processing

Pre-processing is the first step of the character recognition process. It converts real-time images into unique format and

978-1-6654-1836-2/21/$31.00 ©2021 IEEE

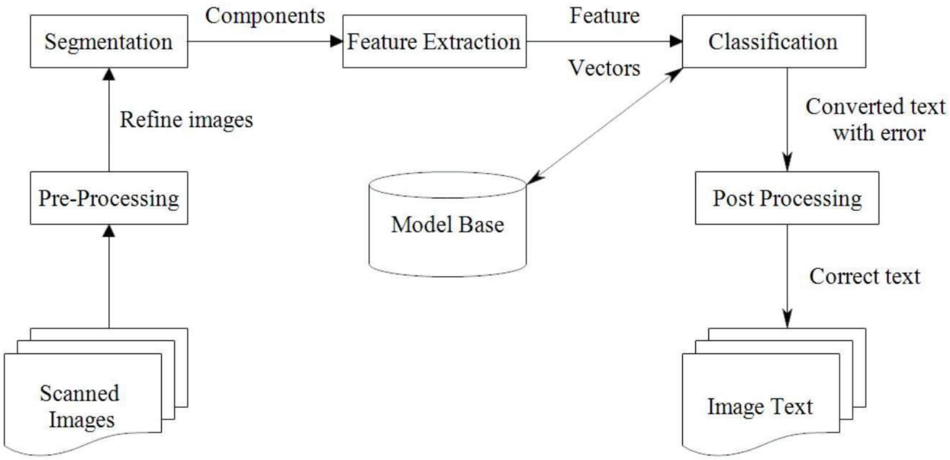

Fig. 1. Working of OCR system [4]

then reduces noise and unwanted background. The methods for pre-processing include noise reduction, binarization, normalization, skew correction, etc. [5]–[8]. As discussed in [6], the further steps, feature extraction step and classification step, depend upon the pre-processing step.

C. Segmentation

The pre-processing step normalizes input images. In the segmentation step, words and characters are retrieved from images. The goal of segmentation is to obtain partitioning of an image into a set of disjoint regions that are different and meaningful with respect to different characteristics.

D. Feature Extraction

Prior to classification, feature extraction is an important step in OCR (Fig. 1). The extracted features have a big impact on OCR recognition accuracy. In this step, for all input images, specific features (characteristics) are extracted and stored into a feature vector. These features are divided into (1) structural features and, (2) statistical features [9].

Statistical features are either local or global features. Global features are extracted from whole character image whereas local features are extracted from local neighbourhood of the images. Moments, distances, projection, crossing, n-tuples, histogram, and zoning are examples of the most typical statistical features. Furthermore, noise and distortions have no impact on global or local features [9].

Structural features are based on topological features that reflect both global and local properties of images. Topological features are used to describe object elements, structure, and properties. Typical examples of topological features are loops, end-points, extreme points, intersection, etc. [9].

E. Classification

As shown in Fig. 1, the next step in the working of OCR system is the classification step wherein character is first recognized. Subsequently, the recognized character is classified into a specific predefined class.

TABLE I
VOWEL CHARACTERS

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः

TABLE II
NUMBER CHARACTERS

०	१	२	३	४	५	६	७	८	९

F. Post Processing

The final step of the OCR system is the post processing step. The main goal of this step is to print the corresponding recognized character in the structured text form [10].

III. HINDI SCRIPT AND ITS PROPERTIES

Hindi script consists of 12 vowels, 36 consonants, and 10 numeral characters [8], [11]–[14]. Vowel is written as either an independent character or a combined character generated after applying various Mantras. Characters that are derived by a combination of vowels and consonants are called Barakhadi character. In Hindi, there is a horizontal line at the upper part of every character that is called Shiro Rekha or headline [11]–[13].

Hindi is written from left to right [15]. It has no upper-case and lower-case characters as in English language. Moreover, the Hindi character set contains more symbols than that of English. Most of the characters in Hindi script are formed by curves, holes, and also strokes. The words in Hindi language are formed by a sequence of characters joined with the help of the Shiro Rekha. The Hindi character set is shown in Table I-IV.

IV. RELATED WORK

In [11], [12], the authors consider the database comprising of Devanagari archaic manuscripts from varied museums and

TABLE III
CONSONANT CHARACTERS

क	ख	ग	घ	ङ	च
छ	ज	झ	ञ	ट	ठ
ड	ढ	ण	त	थ	द
ध	न	प	फ	ब	भ
म	य	र	ल	व	श
ष	स	ह	क्ष	त्र	ज्ञ

TABLE IV
BARAKHADI CHARACTER FOR 'क'

क	का	कि	की	कु	कू	के	कै	को	कौ	कं	कः

libraries. They normalize the characters into 64x64 pixels character by using the nearest-neighbour interpolation method. In [11], for feature extraction, Discrete Cosine Transform (DCT) zigzag is used by the authors. Decision trees [16], Support Vector Machines (SVM) [16], and Naive Bayes [16] classifiers are used for the recognition of the character from the manuscripts. In addition, the authors use ensemble methods - Ada boost and bagging with the base classifier for accuracy improvement. The maximum accuracy is gained for adaptive boosting with Radial Basis Function (RBF) kernel SVM. Different performance evaluation parameters are used by the authors to assess the quality of the ensemble process. In [12], the authors use a Discrete Cosine Transform (DCT) zigzag and Histogram of Oriented Gradients (HOG) [17], [18] for feature extraction. The authors use Support Vector Machines, Naive Bayes, and Decision trees classifiers for the character recognition from the Devanagari archaic manuscripts.

In [19], the authors use various techniques for feature extraction such as Discrete Cosine Transformation, zoning, gradient features and varied combination of these techniques. Adaptive boosting and bagging is also used for the medieval hand-written Gurmukhi manuscripts recognition improvement. They use Random Forest, SVM, K-NN, Decision trees classifiers and also their combination. They consider medieval hand-written Gurmukhi characters of 43 classes containing 1140 samples.

In [7], the authors propose a supervised classifier approach based on Convolution Neural Network (CNN) and Multi-Layer Perceptron (MLP) for recognition of hand-written Gujarati character. They use scanning, resizing, noise removal, binarization, and skew correction as pre-processing steps and a dataset consisting of 10,000 samples of 59 classes.

In [14], the authors prepare and consider a new dataset for Devanagari script - Devanagari Hand-written Character Dataset. This dataset consists of 92,000 images of 46 different classes of the character of Devanagari scripts. They apply pre-processing steps - resizing of images, padding, converting an image into a grayscale form. They use CNN classifier for recognition of the character.

In the work of [17], the authors consider the database consisting of hand-written Hindi characters. It consists of total 4,428 samples and 108 samples for each character with an aim to ensure different orientations and sizes. They use Profile Projection Histogram (PPH) and Histogram of Oriented Gradient (HOG) for feature extraction. They use multiple classifiers - K-NN, Ensemble Subspace Discriminant, Quadratic SVM, bagged trees, weighted K-NN.

In [8], the authors use the region-based k-means clustering for feature extraction of the character. They use binarization of the image. The separation of character is performed in the pre-processing step. They use a Hindi characters database comprising of 430 samples. They further use SVM and Euclidean distance for classification.

In the work of [20], the author uses the database consists of 10 sample images of each 62 hand-written Devanagari characters. The size of each image is 80x40 pixels. For feature extraction, they use a Histogram of Oriented Gradient (HOG). They use an Artificial Neural Network (ANN) for characters' classification.

In [15], the authors use digital curvelet transform and K-Nearest Neighbour (KNN) classifier to propose a novel approach for recognition of Hindi character. First, their approach segments input images. Then, using curvelet transform, features are extracted by calculating thin and thick image. They consider the database containing 200 images of the character set.

In [21], the authors propose a model for recognition of hand-written numerals in Indic scripts. The model utilizes CNN with backpropagation for error reduction and dropout for data over-fitting. They consider the dataset consisting of three languages - Bangla hand-written numerical dataset, Urdu hand-written numerical dataset, and Hindi hand-written numerical dataset.

V. PARAMETRIC EVALUATION

The parametric evaluation of existing techniques for hand-written character recognition is summarized in Table V. For the purpose, we use the parameters - pre-processing techniques, feature extraction techniques, classification techniques, languages considered, performance measures, number of samples used, number of classes, and types of characters (consonants, vowels, numbers) considered.

VI. CONCLUSIONS

In this paper, we present a comprehensive survey on the character and numeral recognition work carried out for Hand-written Hindi scripts. The paper presents different techniques used to recognize hand-written characters of Hindi script from images. It also summarizes the most commonly used feature extraction techniques and pre-processing techniques. Additionally, this paper provides the parametric evaluation of these existing techniques useful to the researchers in the field of hand-written Hindi character recognition.

References

[1] N. B. Muppalaneni, “Handwritten telugu compound character prediction using convolutional neural network,” in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE, 2020, pp. 1–4, doi: https://doi.org/10.1109/ic-ETITE47903.2020.349.

[2] S. P. Ramteke, A. A. Gurjar, and D. S. Deshmukh, “A novel weighted svm classifier based on sca for handwritten marathi character recognition,” *IETE Journal of Research*, pp. 1–13, 2019, doi: https://doi.org/10.1080/03772063.2019.1623093.

[3] G. Nagy, “Twenty years of document image analysis in pami,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 38–62, 2000, doi: https://doi.org/10.1109/34.824820.

[4] M. Meshesha and C. Jawahar, “Optical character recognition of amharic documents,” *African Journal of Information & Communication Technology*, vol. 3, no. 2, 2007, doi: https://doi.org/10.5130/ajict.v3i2.543.

[5] V. C. Hallur and R. Hegadi, “Handwritten kannada numerals recognition using deep learning convolution neural network (dcnn) classifier,” *CSI Transactions on ICT*, vol. 8, pp. 295–309, 2020, doi: https://doi.org/10.1007/s40012-020-00273-9.

[6] S. D. Prasad and Y. Kanduri, “Telugu handwritten character recognition using adaptive and static zoning methods,” in *2016 IEEE Students Technology Symposium (TechSym)*. IEEE, 2016, pp. 299–304, doi: https://doi.org/10.1109/TechSym.2016.7872700.

[7] J. Pareek, D. Singhania, R. R. Kumari, and S. Purohit, “Gujarati handwritten character recognition from text images,” *Procedia Computer Science*, vol. 171, pp. 514–523, 2020, doi: https://doi.org/10.1016/j.procs.2020.04.055.

[8] A. Gaur and S. Yadav, “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*. IEEE, 2015, pp. 65–70, doi: https://doi.org/10.1109/ETTLIS.2015.7048173.

[9] M. Yadav, R. K. Purwar, and M. Mittal, “Handwritten hindi character recognition: a review,” *IET Image Processing*, vol. 12, no. 11, pp. 1919–1933, 2018, doi: https://doi.org/10.1049/iet-ipr.2017.0184.

[10] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: A review,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000, doi: https://doi.org/10.1109/34.824819.

[11] S. R. Narang, M. K. Jindal, and M. Kumar, “Devanagari ancient character recognition using dct features with adaptive boosting and bootstrap aggregating,” *Soft Computing*, vol. 23, no. 24, pp. 13 603–13 614, 2019, doi: https://doi.org/10.1007/s00500-019-03897-5.

[12] S. R. Narang, M. K. Jindal, and P. Sharma, “Devanagari ancient character recognition using hog and dct features,” in *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. IEEE, 2018, pp. 215–220, doi: https://doi.org/10.1109/PDGC.2018.8745903.

[13] P. M. Kamble and R. S. Hegadi, “Handwritten marathi character recognition using r-hog feature,” *Procedia Computer Science*, vol. 45, pp. 266–274, 2015, doi: https://doi.org/10.1016/j.procs.2015.03.137.

[14] S. Acharya, A. K. Pant, and P. K. Gyawali, “Deep learning based large scale handwritten devanagari character recognition,” in *2015 9th International conference on software, knowledge, information management and applications (SKIMA)*. IEEE, 2015, pp. 1–6, doi: https://doi.org/10.1109/SKIMA.2015.7400041.

[15] G. K. Verma, S. Prasad, and P. Kumar, “Handwritten hindi character recognition using curvelet transform,” in *International Conference on Information Systems for Indian Languages*. Springer, 2011, pp. 224–227, doi: https://doi.org/10.1007/978-3-642-19403-0_37.

[16] A. Dey, “Machine learning algorithms: a review,” *International Journal of Computer Science and Information Technologies*, vol. 7, no. 3, pp. 1174–1179, 2016.

[17] M. Yadav and R. Purwar, “Hindi handwritten character recognition using multiple classifiers,” in *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*. IEEE, 2017, pp. 149–154, doi: https://doi.org/10.1109/CONFLUENCE.2017.7943140.

[18] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, vol. 1. IEEE, 2005, pp. 886–893, doi: https://doi.org/10.1109/CVPR.2005.177.

[19] M. Kumar, S. R. Jindal, M. K. Jindal, and G. S. Lehal, “Improved recognition results of medieval handwritten gurmukhi manuscripts using boosting and bagging methodologies,” *Neural Processing Letters*, vol. 50, no. 1, pp. 43–56, 2019, doi: https://doi.org/10.1007/s11063-018-9913-6.

[20] N. Singh, “An efficient approach for handwritten devanagari character recognition based on artificial neural network,” in *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2018, pp. 894–897, doi: https://doi.org/10.1109/SPIN.2018.8474282.

[21] A. K. Tushar, A. Ashiquzzaman, A. Afrin, and M. R. Islam, “A novel transfer learning approach upon hindi, arabic, and bangla numerals using convolutional neural networks,” in *Computational Vision and Bio Inspired Computing*. Springer, 2018, pp. 972–981, doi: https://doi.org/10.1007/978-3-319-71767-8_83.

TABLE V
PARAMETRIC EVALUATION OF EXISTING TECHNIQUES FOR HAND-WRITTEN CHARACTER RECOGNITION

Sr. No.	**Technique**	**Pre-processing technique(s)**	**Feature extraction technique(s)**	**Classification technique(s)**	**Language(s) considered**	**Performance measure used**	**Number of samples used**	**Number of classes considered**	**Types of characters considered**
1	Narang et al. [11]	noise removal, resize of the image	DCT zigzag	decision tree, naive bayes, SVM	Devanagari	Accuracy	5,458	33	consonants
2	Narang et al. [12]	noise removal, resize of the image	DCT zigzag, HOG	decision tree, naive bayes, SVM	Devanagari	Accuracy	5,484	33	consonants
3	Kumar et al. [19]	scanning, resize of the image, normalization	Gradient features, DCT, Zoning	Random Forest, SVM, K-NN, Decision tree	Gurmukhi	Accuracy	1,140	43	vowels and consonants
4	Pareek et al. [7]	noise removal, binarization, skew correction	–	CNN, MLP	Gujarati	Accuracy	10,000	59	vowels and consonants
5	Acharya et al. [14]	resize of the image	CNN	CNN	Hindi	Accuracy	92,000	46	consonants and numbers
6	Yadav et al. [17]	thinning, noise removal	HOG, PPH	SVM, K-NN, bagged trees	Hindi	Accuracy	4,428	41	vowels and consonants
7	Guar et al. [8]	noise removal, binarization	K-means clustering	SVM	Hindi	Accuracy	430	33	consonants
8	Singh [20]	resize of the image, binarization	HOG	ANN	Hindi	Accuracy	400	12	vowels
9	Verma et al. [15]	normalization, noise removal	Unequispaced Fast Fourier Transform	K-NN	Hindi	Accuracy	200	49	vowels and consonants
10	Tushar et al. [21]	–	CNN	CNN	Hindi, Urdu, Bangla	Accuracy	12,000	10	numbers