EasyChair Preprint

№ 2128

A Review on Handwritten Devanagari Character Recognition

Manoj Sonkusare, Roopam Gupta and Asmita Moghe

April 17, 2020

A Review on Handwritten Devanagari Character Recognition

Manoj Sonkusare[1] , Roopam Gupta[2], Asmita Moghe[3]

[1]Research Scholar, Department of Information Technology,
RGPV, Bhopal, India
[2]Professor, Department of Information Technology,
RGPV, Bhopal, India
[3]Professor, Department of Information Technology,
RGPV, Bhopal, India
sonkusare_manoj@rediffmail.com, roopamgupta@rgtu.net,
aamoghe@rgtu.net

Abstract— Because of the vast variation in writing styles, the handwritten text recognition is a challenging task. In India, a large number of people use Devanagari Script to write their documents, but due to large complexity, research work accomplished on this script is much lesser as compared to English script. Hence, recognition of handwritten Devanagari Script is amongst the most demanding research areas in the field of pattern recognition. Feature extraction and classification are important steps of OCR which affects the overall accuracy of the character recognition system. This paper gives a comparative study on different techniques used for feature extraction and classification by the researchers over the last few years.

Keywords: OCR, Devanagari, Handwritten character recognition, online, offline, ANN, K-NN, SVM, CNN.

1 Introduction

Optical Character Recognition (OCR) is a technique for converting the scanned images of handwritten or machine-printed text into a digital form. Handwritten character recognition a active fields of research which possesses a great importance in digital image processing. It has many applications like automation of various organizations like post offices, government and private offices, searching data from documents and books, processing of cheque in banks, etc. OCR is categorized into two

parts which are known as handwritten character recognition and printed character recognition [1], [2]. Recognition of printed text is almost a resolved task. However, due to vast variation in writing styles, the handwritten text recognition is taken into account to be a hard task. Hence, handwritten character recognition is presently an active field of research [3].

The main objective of recognizer is to recognize an image document by splitting up the document into lines and then these lines are further broken into words and in the end into characters [4]. The features are then extracted from these individual characters and compared with the image patterns in order to recognize characters. Deep Neural Networks on the other hand do not require any feature to be explicitly defined, instead they work on the raw pixel data generating the best features at lower layers and using them to classify the inputs into different classes at the upper layer [33], [48]. The recognition of handwritten characters can be carried out either online or offline. In case of online, the current information is available like coordinates of pens tip as a function of time however, in case of offline recognition [5], the image of handwritten paper is required in digital form as shown in figure 1 below:

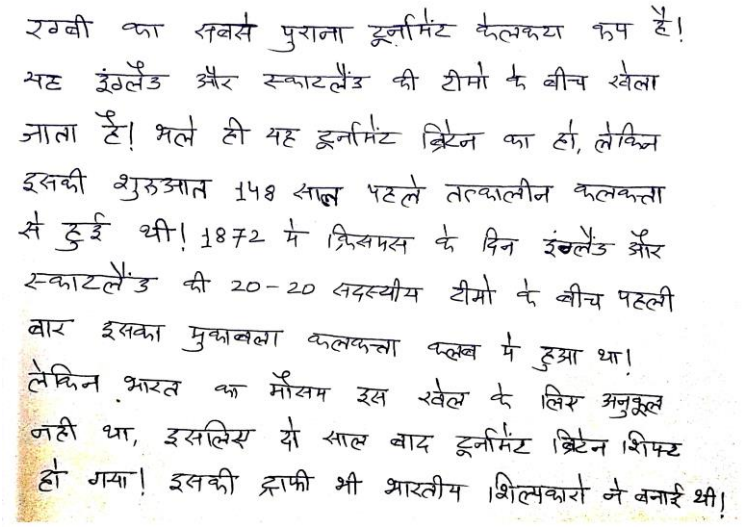

Fig.1. Contemporary handwriting in Devanagari Script.

However, due to gradual progressive growth of handwritten Devanagari character recognition, it is presently new and challenging area. It is developed with the concepts of artificial intelligence, machine learning, pattern recognition, and data mining. Although many handwritten Devanagari character recognition methods have already

been introduced till date, but it is still a complex task to process its documents because of large character set, linguistic based criticalities, and use of shirorekha [6].

Almost all of the classification techniques in the OCR deals with a numerous number of classes and finds discrimination between classes. There are numerous classification techniques available for the problem of pattern recognition. These techniques include, SVM, BPNN, ANN, KNN, CNN and Hybrid Classifier [7]-[9].

The objective of this paper is to find out most accurate feature extraction and classification techniques for recognition of Off-line Handwritten Devanagari script, used by the researchers over the past few years. The intention of this review paper is to serve as a guide for the researchers, working in the similar field.

The paper is managed in the following way: Section II illustrates some information and complexity about Devanagari script. In Section III, we discuss about four different phases, namely: preprocessing, segmentation, feature extraction and character recognition phase. Section IV shows the literature review on Devanagari script. Finally, in last section we discuss conclusion and future scope of the research.

2 Introduction and Complexity of the Devanagari Script

Every Indian language is derived from ancient Brahmi script and has a phonetic base. India has 10 major scripts namely, Devanagari, Kannada, Bangla, Telugu, Gurumukhi, Tamil, Oriya, Malayalam, Urdu and Gujarati. From these scripts, many official languages are derived. Approximately half of the Indian people use Devanagari script and it is used in more than 100 languages like, Hindi, Nepali, Marathi, Haryanvi, Rajasthani, Gujarati, Kashmiri, Bhojpuri, and Sanskrit, etc [10]. It has 49 primary characters from which 13 are vowels, 36 are consonants and 14 are modifiers as shown in figure 2.

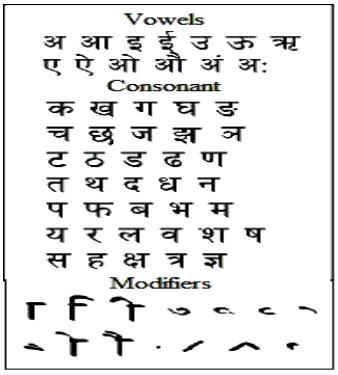

Fig.2. Vowels (Swar), Consonants (Vyanjan) and modifiers of Devanagari script.

"Shirorekha" or "Matra" is also present in Devanagari script. Devanagari words are divided into 3 parts: a core strip that contains main characters, a lower modifier symbol strip (Bottom strip), and an upper modifier symbol strip (Top strip) as shown in figure 3.

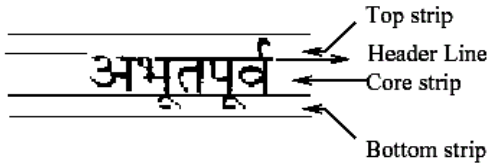

Fig.3. Three strips of Devanagari word

Following are some major points about the complexity of Devanagari script [12].

2.1 Shirorekha

In a word, multiple characters and modifiers appear as a single component connected through the common Shirorekha. Along with numerals, component characters, consonants, vowels, and vowel modifiers, there are many similar shaped characters existing in Devanagari script. All of these collectively make the handwritten character recognition, a difficult task.

2.2 Compound Characters

Compound character joins two or more characters in various ways, forming a new character. Due to structural complexity, recognition of compound characters is a challenging task. Optical Character Recognition for Devanagari script is highly complex due to its large set of conjuncts.

2.3 Size, shape and style variation

In spite of a number of good recognition algorithms that have been existing, the achieved performance is still unreachable from human beings due to the changeable writing conditions and different handwriting styles. There is wide variety of style, shape and size of each character written by writer.

3 Steps in the Recognition Process

Number of factors like noise which is introduced at the time of scanning, broken lines, broken words, or broken characters, various font sizes, etc. affects the result of recognition system. The four different phases in OCR are shown in figure 4 sequentially:

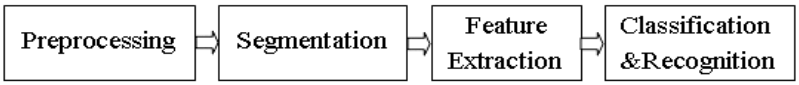

Fig.4. Different phases in optical character recognition system

3.1 Preprocessing

Preprocessing is the first and essential stage for any recognition system to achieve good recognition rate. In this stage, smoothing, enhancing, and filtering techniques are applied to improve readability of digital image. Subsequent algorithm of OCR software uses this digital image for further processing. The various preprocessing stages are shown in figure 5.

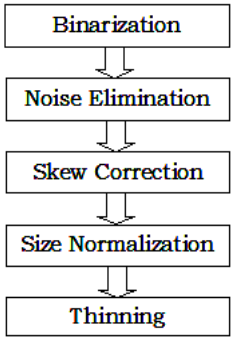

Fig.5. preprocessing stages

3.1.1 Binarization

It is the process of converting a gray scale image to binary image by using thresholding. Two intensity values are obtained as Black & White through this process. The two methods used for conversion of gray level image to binary form are local or adaptive threshold and global threshold. Local or adaptive threshold uses

different pixel values according to the information from local area. Single threshold value is selected in global threshold based on estimation of the background value based on intensity histogram of the image. The binarization process is used to identify the objects [13], [14].

3.1.2 Noise elimination

In scanned image, there may be possibility of noise such as distortion, incomplete corners and gap in the lines. The accuracy in any recognition system may be reduced due to presence of noise. For noise elimination, many morphological and filtering techniques can be applied [13].

3.1.3 Skew Correction

Skewed characters or lines may reduce the accuracy of the succeeding processes, such as segmentation and classification. Hence, it is necessary to make skewed lines horizontal by calculating skew angle and making systematic correction in the scanned image [15]-[17].

3.1.4 Size normalization

The handwritten characters of different writers are not uniform in size. Generally, input provided to the recognition system is an array of definite size. So, to make the image appropriate for recognition system, size normalization is necessary. This process changes the size of image without changing the structure of image [13].

3.1.5 Thinning

A morphological operation called as thinning is applied to remove some unused foreground pixels from binary images. This operation is also called as skeletonization [14].

3.2 Segmentation

Segmentation is also an important stage that affects the accuracy of the character recognition. It is the process of division of a handwritten character image into different characters. Segmentation is obtained by finding the boundaries between characters. There are several techniques for finding the boundaries between characters [18]-[21].

3.3 Feature Extraction

This is the process in which most representative information is extracted from the raw data. For each class, a set of features is extracted that helps to reduce the pattern

variation within class while maximizing it between different classes [22]-[24]. Some feature extraction methods are as follows:

1) Fourier Transforms
2) Gabor Transform
3) Wavelets
4) Moments
5) Zoning
6) Crossings and Distances
7) Projections
8) Coding
9) Graphs and Trees

Deep Neural Networks do not require any feature to be explicitly defined instead they work on the raw pixel data, generating the best features and using them to classify the inputs into different classes [46]. In Deep Convolution Neural Network (DCNN), the features are computed at the lower layers and classification takes place at the uppermost part of the network [33], [48].

3.4 Character Classification and Recognition

Features from input images are extracted and given as an input to the trained classifier like SVM and ANN. Trained Classifiers find out the best matching class by comparing the stored pattern with input features. Various methods and comparative study of Devanagari OCR can be found in [7], [13], [25]. Finally recognized image is converted into editable text. The mostly used character classification techniques are as follows:

1. Artificial Neural Network
2. Support Vector Machine
3. K-nearest neighbors
4. Convolution Neural Networks
5. Hybrid Network

4. Literature Review

In this section we discuss the detailed review on feature extraction and classification techniques used for Devanagari script:

The basic concepts of pattern recognition, understanding of various research models and related algorithms for classification and clustering are introduced in [8]. The paper presents the algorithms for the classification, regression, clustering, parsing and sequence labeling on pattern recognition.

There are three steps in segmentation: Line detection and segmentation, Word detection and segmentation and Character detection and segmentation. In [26], projection profile method is used for segmentation. The horizontal projection profile method is used for line segmentation, vertical projection profile method is used for word segmentation and both horizontal and vertical projection profile method is used to separate characters from words. In [27], new segmentation techniques are suggested on the basis of structure technique for Handwritten Hindi text. In which handwritten text is divided into lines then the lines into words and words into characters. The evaluation of performance on handwritten data of 1380 words of 200 lines is written by 15 different writers. The overall results of segmentation are very efficient.

The template based and feature-based approach was used in older Devanagari character recognition. In the template-based approach, each unknown letter is matched with a standard template pattern of a character and similarity between the two patterns is used to decide which letter it is. Results of the early OCRs were further improved by making use of feature-based approaches along with the traditional template-based approach [28], [29]. Feature-based approaches identify unique characteristics of letters and use them for classification.

The performance of recognition system is improved using principal component analysis (PCA) and Linear Discriminant Analysis (LDA) in [30]. In this work, first of all, raw features are extracted using chain coding, edge detection and direction feature techniques, which are then reduced by LDA and PCA. The SVM classifier is used for characters classification. The state of the art of handwritten and machine printed Devanagari OCR techniques such as training, feature extraction, classification and matching are discussed in various sections of paper [31]. An effort is made to report the most important results. A comparative study and survey on handwritten character recognition for Devanagari script with respect to features and classifiers is presented in [32]. Four feature extraction techniques namely, chain code histogram, shadow, intersection and straight line fitting features are used in [34]. Chain code histogram features, intersection features, and line fitting features are computed by dividing the character image into different segments while shadow features are computed globally for character image. The overall recognition rate of 92.80% was obtained for Devanagari characters with a dataset of 4900 samples.

The recognition of handwritten Devanagari characters using convolution neural networks (CNN) is present in [33]. The architecture consisting of 7 convolution layers and 2 fully connected layers has obtained highest recognition accuracy of 96.09%. Authors discussed, Euclidean distance based KNN techniques for feature extraction, which achieved higher recognition rate than SVM [35]. In [36], curvatures and gradient features are extracted and applied on Euclidean distance Neighbor based Kohonen NN. In [37], features are detected from lines and curves by using Hough

transform and SVM classifier is used for classification. The recognition accuracy obtained from these two methods is up to 90%.

In [38], features like chain code, shadow, histogram, longest run and view based are extracted and fed to NN and in SVM for classification. They achieved 82.89 percent accuracies on Devanagari character databases. The method proposed in [39] uses 16D rotation invariant Ring and Zernike features for recognition of Devanagari handwritten numerals. In [40], [41] constellation diagram and chords structures are used to compute features from characters. In [42], statistical features are evaluated on different sub bands of Haar wavelet transform for recognition of Devanagari characters. Fuzzy technique is used in [43], for recognition of handwritten Hindi characters and numerals and obtained accuracy 90.65 percent and 92.67 percent for handwritten Devanagri characters and numerals respectively. In [44], Hindi OCR is developed by removal of shirorekha in preprocessing stage and features are extracted through K - means clustering and linear kernel based technique is used for classification.

An OCR system is presents in [6], which classified Shirorekha-Less and modified Shirorekha-Less characters of handwritten Sanskrit, Hindi and Marathi image documents using support vector machine. The system was developed on various datasets of these languages and achieved better result of 98.35%. A robust algorithm for character segmentation and recognition for Latin and Devanagari scripts is proposed in [11]. In this work, primary segmentation paths are obtained by using structural property, whereas overlapped and joined characters are segmented using graph distance theory and then SVM classifier validates the segmentation results. KNN classifier is used for recognizing the handwritten and printed input characters and obtained recognition rates of 97.05% for Devanagari script and of 97.10% for Latin script respectively. An array based feature extraction and back propagation Neural Network for recognition of handwritten Marathi characters is proposed in [45]. The experiments are conducted on 500 handwritten characters obtained from 10 different persons. The recognition accuracy obtained is 92%. An analysis of pre-trained models using transfer learning for Deep Convolution Neural Network (DCNN) is present in [47]. Convolutional layers, pooling layers and fully connected layers of CNN are used for feature extraction, dimensionality reduction and image representation. In this work 15 epochs are implemented for each of 7 pre-trained models. The recognition results show maximum accuracy of 99% for handwritten Devanagari alphabets. Two deep learning models are used for recognition and to train the dataset in [48]. This work also analyzes the effect of dataset increment and dropping out units approach to prevent over-fitting of the networks. The experimental results suggest that Deep CNNs with dataset increment technique and added Dropout approach can result in test accuracy of 98.47%.

Table 1. A Comparative Study of five recent works on off-line handwritten character recognition using different classifiers

S. No	Work	Feature Extraction	Classifier	Accuracy %
1	Nagender Aneja et. al.[47]	These networks do not require any feature to be explicitly defined, instead they work on the raw pixel data generating the best features at lower layers and using them to classify the inputs into different classes at the upper layer of the network.	CNN	99%
2	Shailesh Acharya et. al.[48]	-Do-	CNN	98.47%
3	Shalini Puri et. al.[6]	Geometric based features	SVM	98.35%
4	Parul sahare et. al.[11]	Fixed center distance based feature, Fixed center cut based feature, Neighborhood counts based feature	Hybrid Classifier (SVM, K-NN)	97.05%
5	Pankaj kale et. al.[45]	Array based feature extraction	ANN	92%

5. Conclusion and Future Scope

In this paper, we present detailed study on off-line handwritten Devanagari character recognition and comparison between five recent works with highest recognition accuracy. The comparison of recent work using different classifiers is shown in Table 1. Various feature extraction and character recognition techniques used are also discussed in this survey. The survey concludes that CNN provides better results than traditional networks with the recognition accuracy of 99%. This study points out that the work done on Devanagari scripts is still at its initial stage, so it needs further research to solve many problems. The future scope of the work area is as follows:

- It can be extended to recognize words, sentences and the characters in actual world images.
- Various Indian and Latin scripts can be included in future work to make it a generic system.

- The dataset of the system contains only Vowels and Consonants. It can be extended with numerals and then used for Devanagari Characters as well as Numeral Recognition.

References

1. Arica, N., and Vural F. Y.: An Overview of Character Recognition Focused on Off-line Handwriting. IEEE Transactions on Systems man and cybernetics, vol. 31, pp. 216 – 233, (2000).
2. Kumar M., Jindal M. K., and Sharma R. K.: Review on OCR for Handwritten Indian Scripts Character Recognition. DPPR 2011, CCIS 205, pp. 268–276, (2011).
3. Bhunia A. K., Roy P. P., Mohta A., and Pal U.: Cross-language framework for word recognition and spotting of indic. script. Pattern Recognition, Vol. 79, (2018).
4. Jindal K., and Kumar R.: A new method for segmentation of pre-detected Devanagri words from the scene images: Pihu method. Computers & Electrical Engineering, Vol. 70, pp. 754-763, (2018).
5. Singh P. K., Sarkar R., and Nasipuri M.: Offline script identification from multilingual Indic-script documents: A State-of-the-art. Computer Science Review, Vol. 15-16, pp. 1-28, (2015).
6. Puri, S., and Singh, S. P.: An efficient Devanagari character classification in printed and handwritten documents using SVM. International Conference on Pervasive Computing Advances and Applications, Vol. 152, pp. 111-121, (2019).
7. Gour R, and Chouhan V. S.: Classifiers in Image processing. International Journal on Future Revolution in Computer Science & Communication Engineering, Vol. 03, pp. 22-24, (2017).
8. Sharma P., and Kaur M.: Classification in Pattern Recognition: A Review. International Journal of Advance Research in Computer Science & Software Engineering, Vol. 03, pp. 298-306, (2013).
9. Mane D. T., and Kulkarni U. V.: Visualizing and Understanding Customized Convolutional Neural Network for Recognition of Handwritten Marathi Numerals. International Conference on Computational Intelligence and Data Science, pp.1123-1137, (2018).
10. Sharma R., Kaushik B. N., and Gondhi N. K.: Devanagari and Gurmukhi Script Recognition in the Context of Machine Learning Classifiers : Mini Review. Journal of Artificial Intelligence, pp. 65-70, (2018).

11. Sahare P., and Dhok S. B.: Multilingual character Segmentation and Recognition schemes for Indian document images. IEEE Trans., Vol. 06, pp 10603-10617, (2018).
12. Ajmire P. E., Dharaskar R. V., and Thakare V. M.: Study of Complexity of Devanagari Script for Handwritten Character Recognition. Proceedings of Twelveth IRF International Conference, (2014).
13. Bhopi S. A., and Singh M. P.: Review on Optical Character Recognition of Devanagari Script Using Neural Network. International Journal on Future Revolution in Computer Science & Communication Engineering, Vol. 04, pp. 415-420, (2018) .
14. Indira B., Qureshi M. S., Shaik M. S., and Saqib S. M.: Devanagari Character Recognition: A Short Review. International Journal of Computer Applications, Vol. 59, pp. 23-27, (2012).
15. Cheriet M., Kharma N., Liu C. L., and Suen C. Y.: Character Recognition Systems: A Guide for students and Practioners. John Wiley & Sons, Inc., Hoboken, New Jersey, (2007).
16. Kapoor R., Bagai D., and Kamal T. S.: Skew angle detection of a cursive handwritten Devnagari script character image. Journal of Indian Inst. Science, pp. 161–175, (2002).
17. Pal U., Mitra M. and Chaudhuri B. B.: Multi-Skew Detection of Indian Script Documents. pp. 292-296, (2001).
18. Shukla M. K., and Banka H.: An Efficient Segmentation Scheme for the Recognition of Printed Devanagari Script. International Journal of Computer Science and Technology, Vol. 02, pp. 529-531, (2011).
19. Shukla M. K., Patnaik T., Tiwari S., and Singh S. K.: Script Segmentation of Printed Devanagari and Bangla Language Document images OCR. International Journal of Computer Science and Technology, Vol. 02, pp. 367-370, (2011).
20. Indira B. and Sudha T.: A Pragmatic Approach for Reading Number Plates of Indian Vehicles. International Journal of Neural Networks and Applications, Vol. 03, pp. 15-18, (2010).
21. Casey R. G. and Lecoline E.: A survey of methods and strategies in Character segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, pp. 690-706, (1996).
22. Trier D., Jain A. K., and Taxt T.: Feature Extraction Method for Character Recognition– A Survey. Pattern Recognition, Vol. 29, pp. 641-662, (1996).
23. Oh I. S., Lee J. S., and Suen C. Y.: Analysis of class separation and Combination of Class-Dependent Features for Handwriting Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, pp.1089-1094, (1999).

24. Dongre V. J., and Mankar V. H.: A Review of Research on Devnagari Character Recognition. International Journal of Computer ApplicationsVol. 12, pp. 8-15, (2010).
25. Pal U., Wakabayashi T., and Kimura F.: Comparative Study of Devanagari Handwritten Character Recognition Using Different Features and Classifiers. 10th International. Conference on Document Analysis and Recognition, pp. 1111-1115, (2009).
26. Kant A. J., Vyavahare A. J.: Devanagari OCR Using Projection Profile Segmentation Method. International Research Journal of Engineering and Technology (IRJET), Vol. 03, (2016).
27. Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal, "Segmentation of Handwritten Hindi Text", Volume 1, No. 4, 2010.
28. Pal, U., and Chaudhuri B. B.: Indian script character recognition: a survey. Pattern Recognition, pp.1887-1899, (2004).
29. Bansal, Veena, and Sinha: Integrating knowledge sources in Devanagari text recognition system. IEEE Transactions on Systems, Man, and Cybernetics-Part A, pp. 500-505, (2000).
30. Sithole S. and Jadhav S.: Recognition of Handwritten Devanagari characters using Linear Discriminant Analysis. ICISC 2018, pp. 100-103, (2018).
31. Jayadevan R., Kolhe S. R., Patil P. M., and Pal U.: Offline recognition of Devanagari script: A survey. IEEE Transactions on Systems, Man, and Cybernetics, pp.782–796, (2011).
32. Indian, A., and Bhatia, K.: A survey of offline handwritten Hindi character recognition. Third International Conference on Advances in Computing Communication & Automation, pp. 1–6, (2017).
33. Chakraborty B., Shaw B., Aich J., Bhattacharya U., and Parui S. K.: Does Deeper Network Lead to Better Accuracy: A Case Study on Handwritten Devanagari Characters. 13th IAPR International Workshop on Document Analysis and Systems, (2018).
34. Arora S.: Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition. IEEE Trans., (2008).
35. Holambe A. K. N. , Thool R. C., and Jagade S. M.: A brief review and survey of feature extraction methods for Devnagari OCR. Ninth International Conference on ICT and Knowledge Engineering, pp. 99–104, (2012).
36. Holambe and Thool: Comparative Study of devanagri handwritten & printed character & Numerals recognition using nearest neighbor classifiers. Ninth International Conference on ICT and knowledge Engineering, (2011).
37. Nene A., Palkar A., Nagarhalli M., Misri R., and Kone S.: Survey on handwritten devanagari character recognition. International Education & Research Journal [IERJ], Vol. 03, pp. 45-46, (2017).

38. Arora S.: Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition. IJCSI International Journal of Computer Science Issues, Vol. 7, (2010).
39. Patil P. M., and Sontakke T. R.: Devanagari Numeral and Character Recognition using general fuzzy neural network. Pattern Recognition, pp. 2110–2117, (2007).
40. Dash K. S., Puhan N. B., and Panda G.: Unconstrained handwritten digit recognition using perceptual shape primitives. Pattern Analysis and Applications, pp. 1-24, (2016).
41. Dash K. S., Puhan N. B., and Panda G.: BESAC: Binary external symmetry axis constellation for unconstrained handwritten character recognition. Pattern Recognition , vol. 83, pp. 413-422, (2016).
42. Dixit A., Navghane A., and Dandawate Y.: Handwritten Devanagari character recognition using wavelet based feature extraction and classification scheme. 11th IEEE INDICON, pp. 1–4, (2014).
43. Hanmandalu M., and Murthy O. V. R.: Fuzzy model based recognition for handwritten Hindi numerals. International conference on Recognition, pp. 490-496, (2005).
44. Gaur A., and Yadav S.: Handwritten Hindi character recognition using K-means clustering and SVM. Fourth International Symposium on Emerging Trends and Technologies in Libraries and Information Services, pp. 65–70, (2015).
45. Kale P., Bang A. V., and Joshi D.: Recognition Of Handwritten Devanagari Characters Using Machine Learning Approach. International Journal of Industrial Electronics and Electrical Engineering, Vol. 03, pp. 48-51, (2015).
46. Lee H., Grosse R., Ranganath R., and Ng. A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. 26th Annual International Conference on Machine Learning, pp. 609–616. (2009).
47. Aneja N., and Aneja S.: Transfer Learning using CNN for Handwritten Devanagari Character Recognition. IEEE International Conference on Advances in Information Technology (ICAIT), pp. 1-4, (2019).
48. Acharya S., Pant A. K., and Gyawali P. K.: Deep Learning Based Large Scale Handwritten Devanagari Character Recognition. 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), (2015).