EasyChair Preprint

№ 3668

A Comparative study on Handwritten Devanagari Character Recognition

Manoj Sonkusare, Roopam Gupta and Asmita Moghe

June 22, 2020

A Comparative study on Handwritten Devanagari Character Recognition

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

[1]Research Scholar, [2]Professor, [3]Professor,
[1,2,3] Department of IT, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India
sonkusare_manoj@rediffmail.com

Abstract. Handwritten text recognition is a challenging task because of the vast changes in writing styles. In India, a massive 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 image processing. Feature extraction and recognition are key steps of OCR which affects the accuracy of the character recognition system. This paper gives a comparative study on distinct techniques used for feature extraction and classification by the researchers over the last few years.

Keywords: OCR, Devanagari Script, ANN, CNN, K-NN, SVM.

1 Introduction

Optical Character Recognition (OCR) is a technique to turn the scanned image of handwritten or printed text into a digital form. Handwritten character recognition is active field of research which possesses a substantial importance in digital image processing. It has many applications such as automation of various organizations like post offices, government and private offices, searching data from documents and books, processing of cheque in banks, etc. Handwritten character recognition and printed character recognition are two types of OCR [1], [2]. Printed text recognition is almost a resolved job. However, because of the vast changes in writing styles, the handwritten text recognition is a challenging job. Hence, handwritten character recognition is presently a vital field of research [3].

Either online or offline both types of handwritten characters can be supported by the recognition system. In the former case, the current information is available as coordinates of pens tip as a function of time however, in later case the image of handwritten paper is required in digital form [4] as shown in figure 1 below:

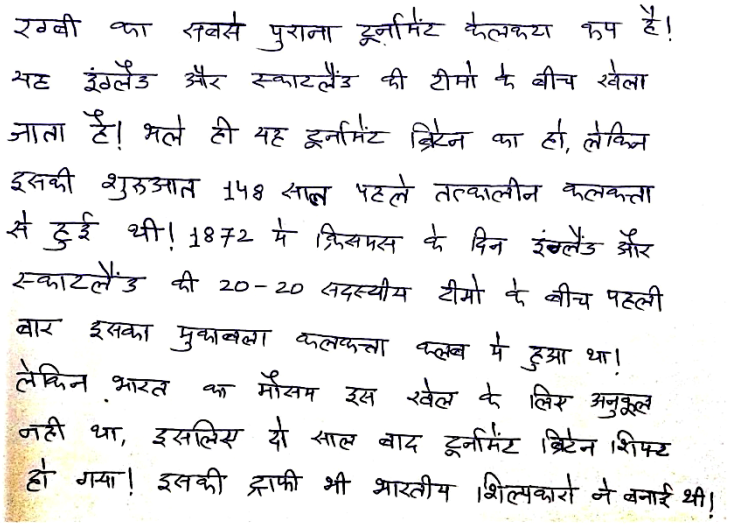

Fig.1. Contemporary handwriting in Devanagari Script.

However, due to gradual progressive growth of handwritten Devanagari character recognition, it is presently new and challenging area. 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 [5].

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 such as, SVM, BPNN, ANN, KNN, CNN and Hybrid Classifier [6]-[8].

The purpose of this paper is to find out most accurate feature extraction and classification techniques used by the researchers over the past few years. The intention of this paper is to direct the researchers who are pursuing their work in the similar field.

The paper is managed in the subsequent manner: Section 2 illustrates basic information about Devanagari script. In Section 3, we elaborate four distinct phases, that is: preprocessing, segmentation, feature extraction and character recognition phase. Section 4 shows the literature review on Devanagari script. Finally, in last section we discuss conclusion and future scope of the research.

2 Introduction to Devanagari Script

Every Indian language is extracted from Brahmi script and has a phonetic base. India has 10 major scripts that is, Gujarati, Kannada, Oriya, Telugu, Gurumukhi, Tamil, Bangla, Malayalam, Urdu and Devanagari. From these scripts, many official languages are extracted. Approximately half of the Indians use Devanagari script and it is used in more than 100 languages like, Hindi, Nepali, Marathi, Haryanvi,

Rajasthani, Gujarati, Kashmiri, Bhojpuri, and Sanskrit, etc [9]. 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, Consonants and modifiers of Devanagari script.

"Shirorekha" or "Matra" is also present in Devanagari script. Devanagari words are divided into 3 parts: a core strip containing base characters, a Bottom strip with lower modifier, and Top strip having upper modifier as displayed in figure 3.

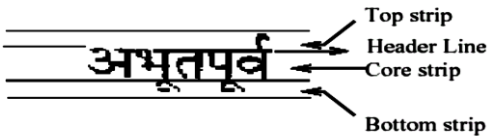

Fig.3. Devanagari word strips

3 Steps in the Recognition Process

The number of factors which is introduced at the time of scanning is broken lines, broken words, or broken characters, affects the result of system. The four distinct steps in OCR are shown in figure 4 sequentially:

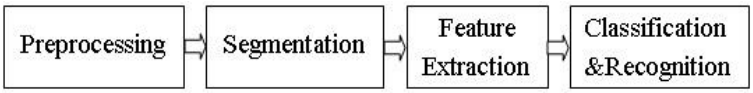

Fig.4. Distinct steps in optical character recognition system

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.

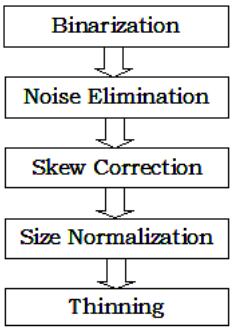

Fig.5. preprocessing steps

Binarization. It is the process to turn from gray scale to binary image in which black and white images are obtained as intensity variations. The two methods used for binarization are local and global threshold. Local threshold uses distinct pixel values based on the local spatial information. Single threshold value is selected in global threshold based on information from the whole image. The aim of binarization is to identify the objects [11], [12].

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 [11].

Skew Correction. Skewed characters or lines may reduce the recognition precision of the succeeding task like segmentation and classification. Hence, it is necessary to make skewed lines horizontal by calculating angle and making systematic alteration in scanned image [13]-[15].

Size Normalization. The handwriting of different writers is not uniform in size. So, in order to produce the image appropriate for recognition system, size normalization is necessary. This process changes the dimension of image without changing the shape of image [11].

Thinning. Thinning is applied to eliminate some unused foreground pixels from binary images. This operation is also called as skeletonization [12].

3.2 Segmentation

Segmentation is also a vital step that affects the accuracy of the system. It is the operation of division of image to different characters. There are diverse techniques to locate the boundaries between characters [16]-[19].

3.3 Feature Extraction

This is the process in which most symbolic information is extracted from the raw data. For each class, features are extracted that helps to reduce the pattern variation within class while maximizing it between different classes [20]-[22]. 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

In Deep Convolution Neural Network (DCNN), the best features are generated from raw data and used to classify the inputs into different classes [38]. In DCNN, the lower layer evaluates the features and upper layer performs classifications [31], [40].

3.3 Character Classification and Recognition

Extraction of features from input images serves as an input to the trained classifier namely SVM, ANN etc. Trained Classifiers locate the best matched class by comparing the stored pattern with input features. Various OCR methods and their comparison can be found in [6], [11], [23]. Finally recognized image is converted into editable text. The mostly used character classification techniques are as follows:

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

4. Literature Review

In this segment we discussed distinct feature extraction and techniques for classification:

The basic pattern recognition ideas, understanding of various research models and related algorithms for classification and clustering are introduced in [7]. The paper presents the algorithms for the classification, regression, clustering, parsing and sequence labeling on pattern recognition.

There are three steps in segmentation: Line, Word and Character detection and segmentation. In [24], horizontal projection profile method for line, vertical projection profile method for word and both horizontal and vertical projection profile method to segment characters from words. In [25], new segmentation techniques are suggested, in which scanned handwritten image is divided into lines, these lines to words and words to characters.

The feature and template-based method was applied for older Devanagari character recognition. In template-based method, each input letter is matched with a standard template pattern of characters and similarity between two patterns is used to decide which letter it is. The results of older OCRs were enhanced by making use of feature-based method in addition to traditional template-based method [26], [27]. Feature-based methods find the unique aspect of letters and these aspects are then used in classification.

The performance of recognition system is improved by using principal component analysis and Linear Discriminant Analysis [28]. In which, chain coding, edge detection and direction feature techniques are used for extraction of raw features which are then reduced by LDA and PCA. The SVM is employed for classification of characters. The new techniques of Devanagari OCR such as training, feature extraction, classification and matching are presented in [29]. A comparative study concerning feature extraction approaches and classifiers on Devanagari OCR is explained in [30].

In [31], Devanagari handwritten characters are recognized by use of convolution neural networks (CNN). The architecture consisting of 7 convolution and two fully connected layers has obtained highest accuracy of 96.10%. Authors discussed, Euclidean distance based KNN techniques for feature extraction, which achieved higher recognition rate than SVM [32]. In [33], curvatures and gradient features are extracted and applied on Euclidean distance Neighbor based Kohonen NN. In [34], features are detected from lines and curves by using Hough transform and classification is accomplished by SVM. The accuracy secured from these two methods is up to 90%.

Fuzzy technique is used in [35], for recognition of Hindi handwritten characters and secured accuracy of 90.65 percent for handwritten Devanagri characters. In [36], Hindi OCR is developed by removal of shirorekha in preprocessing stage, K - means clustering approach is used for feature extraction and linear kernel based technique is used for classification.

An OCR system is presented in [5], which classifies and modifies Shirorekha-Less character 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 strong algorithm in Devanagari and Latin scripts for segmentation and recognition is proposed in [10]. In which, primary segmentation paths are acquired through structural property, joined and overlapped characters are segmented using graph distance and then SVM classifier validates the segmentation results. KNN classifier is utilized for handwritten and printed input characters and obtained recognition rates of 97.05% for Devanagari script and of 97.10% for Latin script respectively. Feature extraction based on array and BPNN for recognition of Marathi handwritten characters is proposed in [37]. The test is performed on 500 handwritten characters obtained from 10 different persons. The recognition accuracy obtained is 92%. A pre-trained model using transfer learning for DCNN is presented in [39]. Convolutional, pooling and fully connected layers of CNN are applied for feature extraction, reduction of dimensions 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 [40]. This work also analyzes the effect of dataset increment and dropping out units approach to prevent over-fitting of the networks. The test results suggest that DCNN with dataset increment technique and added Dropout approach results in accuracy of 98.47%.

Table 1. Comparison between five recent works on Devanagari handwritten character recognition

S. No	Work	Feature Extraction	Classifier	Accuracy %
1	Nagender Aneja et. al.[39]	The best features are generated from raw data and used to classify the inputs into different classes	CNN	99%
2	Shailesh Acharya et. al.[40]	-Do-	CNN	98.47%
3	Shalini Puri et. al.[5]	Geometric based features	SVM	98.35%
4	Parul sahare et. al.[10]	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.[37]	Array based feature extraction	ANN	92%

5. Conclusion and Future Scope

This paper presents detailed study on Devanagari handwritten 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 recognition techniques used are also presented 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 carried out on Devanagari scripts is at preliminary stage, therefore it needs more research to solve many similar problems. The forthcoming scope in this area is as follows:

- It can be expanded for recognition of words, sentences and 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., Vural, F.Y.: An Overview of Character Recognition Focused on Off-line Handwriting. IEEE Transactions on Systems man and cybernetics 31(2), 216 – 233 (2001).
2. Kumar, M., Jindal, M.K., Sharma, R.K.: Review on OCR for Handwritten Indian Scripts Character Recognition. In: Nagamalai, D., Renault, E, Dhanushkodi, M. (Eds.): DPPR 2011, CCIS, vol. 205, pp. 268–276. Springer, Heidelberg 2011.
3. Bhunia, A.K., Roy, P.P., Mohta, A., Pal, U.: Cross-language framework for word recognition and spotting of indic. script. Pattern Recognition 79, 12-31 (2018).
4. Singh, P.K., Sarkar, R., Nasipuri, M.: Offline script identification from multilingual Indic-script documents: A State-of-the-art. Computer Science Review 15-16, 1-28 (2015).
5. Puri, S., Singh, S.P.: An efficient Devanagari character classification in printed and handwritten documents using SVM. In: Bundele, M., Dey, N., Madria, S.K. (eds.): International Conference on Pervasive Computing Advances and Applications 2019, Procedia Computer Science, vol. 152, pp. 111-121, Elsevier, (2019).
6. Gour, R, Chouhan, V.S.: Classifiers in Image processing. International Journal on Future Revolution in Computer Science & Communication Engineering 3(6), 22-24 (2017).
7. Sharma, P., Kaur, M.: Classification in Pattern Recognition: A Review. International Journal of Advance Research in Computer Science & Software Engineering 3(4), 298-306 (2013).
8. Mane, D.T., Kulkarni, U.V.: Visualizing and Understanding Customized Convolutional Neural Network for Recognition of Handwritten Marathi Numerals. In: Singh, V., Asari, V.K., Patel, R.B., Sidike, P. (eds.): International Conference on Computational Intelligence and Data Science, vol. 132, pp.1123-1137, Elsevier, (2018).
9. Sharma, R., Kaushik, B.N., Gondhi, N.K.: Devanagari and Gurmukhi Script Recognition in the Context of Machine Learning Classifiers : Mini Review. Journal of Artificial Intelligence 11(2), 65-70 (2018).
10. Sahare, P., Dhok, S.B.: Multilingual character Segmentation and Recognition schemes for Indian document images. IEEE Trans. 06, 10603-10617 (2018).

11. Bhopi, S.A., Singh, M.P.: Review on Optical Character Recognition of Devanagari Script Using Neural Network. International Journal on Future Revolution in Computer Science & Communication Engineering 04(3), 415-420 (2018).
12. Indira, B., Qureshi, M.S., Shaik, M.S., Saqib, S.M., Murthy, MVR: Devanagari Character Recognition: A Short Review. International Journal of Computer Applications 59(6), 23-27 (2012).
13. Cheriet, M., Kharma, N., Liu, C.L., Suen, C.Y.: Character Recognition Systems: A Guide for students and Practioners. 1st Edition John Wiley & Sons Inc., New Jersey, (2007).
14. Kapoor, R., Bagai, D., Kamal, T.S.: Skew angle detection of a cursive handwritten Devnagari script character image. Journal of Indian Inst. Science 82, 161–175 (2002).
15. Pal, U., Mitra, M., Chaudhuri, B.B.: Multi-Skew Detection of Indian Script Documents. In: Proceedings of Sixth International Conference on Document Analysis and Recognition pp. 292-296, IEEE, USA (2001).
16. Shukla, M.K., Banka, H.: An Efficient Segmentation Scheme for the Recognition of Printed Devanagari Script. International Journal of Computer Science and Technology 02(4), 529-531 (2011).
17. Shukla, M.K., Patnaik, T., Tiwari, S., Singh, S.K.: Script Segmentation of Printed Devanagari and Bangla Language Document images OCR. International Journal of Computer Science and Technology 2(2), 367-370 (2011).
18. Indira, B., Sudha, T.: A Pragmatic Approach for Reading Number Plates of Indian Vehicles. International Journal of Systems and Technologies 2(2), 233-240 (2009).
19. Casey, R.G., Lecoline, E.: A survey of methods and strategies in Character segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(7), 690-706 (1996).
20. Trier, D., Jain, A.K., Taxt, T.: Feature Extraction Method for Character Recognition– A Survey. Pattern Recognition 29(4), 641-662 (1996).
21. Oh, I.S., Lee, J.S., Suen, C.Y.: Analysis of class separation and Combination of Class-Dependent Features for Handwriting Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 21(10), 1089-1094 (1999).
22. Dongre, V.J., Mankar, V.H.: A Review of Research on Devnagari Character Recognition. International Journal of Computer Applications 12(2), 8-15 (2010).
23. Pal, U., Wakabayashi, T., Kimura, F.: Comparative Study of Devanagari Handwritten Character Recognition Using Different Features and Classifiers. In: The Proceedings of 10th International. Conference on Document Analysis and Recognition, pp. 1111-1115, IEEE, Barcelona (2009).
24. Kant, A.J., Vyavahare, A.J.: Devanagari OCR Using Projection Profile Segmentation Method. International Research Journal of Engineering and Technology (IRJET) 03(7), 132-134 (2016).
25. Garg, N.K., Kaur, L., Jindal, M.K.: Segmentation of Handwritten Hindi Text. International Journal of Computer Applications 01(4), 19-23 (2010).
26. Pal, U., Chaudhuri, B.B.: Indian script character recognition: a survey. Pattern Recognition 37(9), 1887-1899 (2004).
27. Bansal, V., Sinha RMK: Integrating knowledge sources in Devanagari text recognition system. IEEE Transactions on Systems, Man, and Cybernetics-Part A 30(4), 500-505 (2000).

28. Sithole, S., Jadhav, S.: Recognition of Handwritten Devanagari characters using Linear Discriminant Analysis. In: Second International Conference on Inventive Systems and Control, pp. 100-103, IEEE, Coimbatore, India (2018).
29. Jayadevan, R., Kolhe, S.R., Patil, P.M., Pal, U.: Offline recognition of Devanagari script: A survey. IEEE Transactions on Systems, Man, and Cybernetics-Part C 41(6), 782–796, (2011).
30. Indian, A., Bhatia, K.: A survey of offline handwritten Hindi character recognition. In: The Proceedings of Third International Conference on Advances in Computing Communication & Automation, pp. 1–6, IEEE, Dehradun, India (2017).
31. Chakraborty, B., Shaw, B., Aich, J., Bhattacharya, U., Parui, S.K.: Does Deeper Network Lead to Better Accuracy: A Case Study on Handwritten Devanagari Characters. In: The Proceedings of 13th IAPR International Workshop on Document Analysis and Systems, pp. 411-416, IEEE, Vienna, Austria (2018).
32. Holambe, A.K.N., Thool, R.C., 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, IEEE, Bangkok, Thailand (2011).
33. Holambe, Thool: Comparative Study of devanagri handwritten & printed character & Numerals recognition using nearest neighbor classifiers. In: The Proceedings of 3rd International Conference on Computer Science and Information Technology, pp. 426-430, IEEE, Chengdu, China (2010).
34. Nene, A., Palkar, A., Nagarhalli, M., Misri, R., Kone, S.: Survey on handwritten devanagari character recognition. International Education & Research Journal [IERJ] 03(3), 45-46 (2017).
35. Hanmandalu, M., Murthy, O.V.R., Madasu, V.K.: Fuzzy Model based recognition of handwritten Hindi characters. In: The Proceedings of 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications (DICTA), pp. 454-461, IEEE, Glenelg, Australia (2007).
36. Gaur, A., Yadav, S.: Handwritten Hindi character recognition using K-means clustering and SVM. In: The Proceedings of Fourth International Symposium on Emerging Trends and Technologies in Libraries and Information Services, pp. 65–70, IEEE, Noida, India (2015).
37. Kale, P., Bang, A.V., Joshi, D.: Recognition Of Handwritten Devanagari Characters Using Machine Learning Approach. International Journal of Industrial Electronics and Electrical Engineering 03(9), 48-51 (2015).
38. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: The Proceedings of 26th Annual International Conference on Machine Learning, pp. 609–616. Montreal, Canada (2009).
39. Aneja, N., Aneja, S.: Transfer Learning using CNN for Handwritten Devanagari Character Recognition. In: The Proceedings of 1st International Conference on Advances in Information Technology (ICAIT), pp. 293-296, IEEE, Chikmagalur, India (2019).
40. Acharya, S., Pant, A.K., Gyawali, P.K.: Deep Learning Based Large Scale Handwritten Devanagari Character Recognition. In: The Proceedings of 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), IEEE, Kathmandu, Nepal (2015).