IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.12402

DEEP LEARNING METHOD USING OCR FOR DEVANAGARI SCRIPT

Ranjeet Ramesh Pawar[1], Mithun Vishnu Mhatre[2]

Information Technology Department, BVIT, Navi Mumbai, India[1]

Computer Technology Department, BVIT, Navi Mumbai, India[2]

Abstract: In the discipline of pattern recognition, optical character recognition is a critical task. Many academics have researched English character recognition extensively, however, in the case ofIndian characters, there has been less investigation. Languages that are difficult to understand, extensive research required. Devanagari is a commonly used Indian script. Individuals from India Devanagari is the foundation for a number of languages. Hindi, Sanskrit,Kashmiri, and Marathi are among the Indian languages and so forth. A review of previous studies is presented in this article. Work on Devanagari character recognition as well as a fewuses for an optical character recognition system.

Character recognition is a research problemthat has been ongoing for many years. In optical character recognition, a procedure of automatically recognizing the optically scanned character images and digitized character images is to be developed into an electronic text document. Devanagari is an Indian script that is a very popular script among millions of people. There are many Indian languages thatare the basis of Devanagari. Those languages are Hindi, Sanskrit, Kashmiri, Marathi, and manymore. English character recognition is mostly studied by researchers and a lot of commercialsystems are used for it. But for Indian languages, the research work is very limited because ofthe complex formation of the language.

Keywords: Include at least 4 keywords or phrases.

I. INTRODUCTION

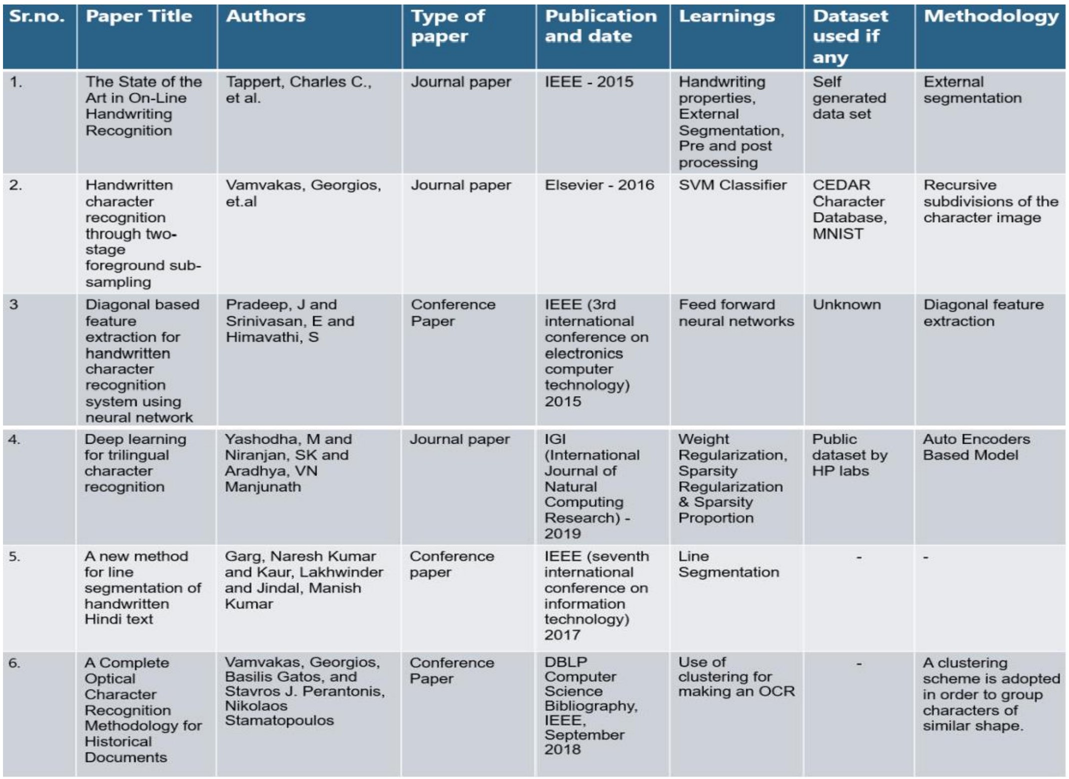
As technology, computing power, and creative sensing and rendering technologiesprogress at a rapid rate, computers are becoming increasingly intelligent. In their study,a number of researchers have proved the capacity of computers to interface or interactwith humans, and a number of commercial products are also available. One suchcapability is optical character recognition (OCR), which entails the automatic conversion of scanned documents/images of machine-printed/handwritten characters into machine-readable digital form. OCR technologies help automate the processing of largevolumes of textual data.

Because many OCR-based approaches can only read/recognize characters from one script, they are script-specific. Because India is a multilingual country with 23 languages and 13 scripts (including English/Roman), automated recognition of both printed and handwritten characters/scripts has a wide range of applications. Because most OCR algorithms are script-specific, recognizing and processing documents containing severalscripts is difficult.

Character recognition is a research problem that has been ongoing for many years. In optical character recognition, a procedure of automatically recognizing the optically scanned character images and digitized character images is to be developed into an electronic text document. Devanagari is an Indian script that is a very popular script among millions of people. There are many Indian languages that are the basis ofDevanagari.

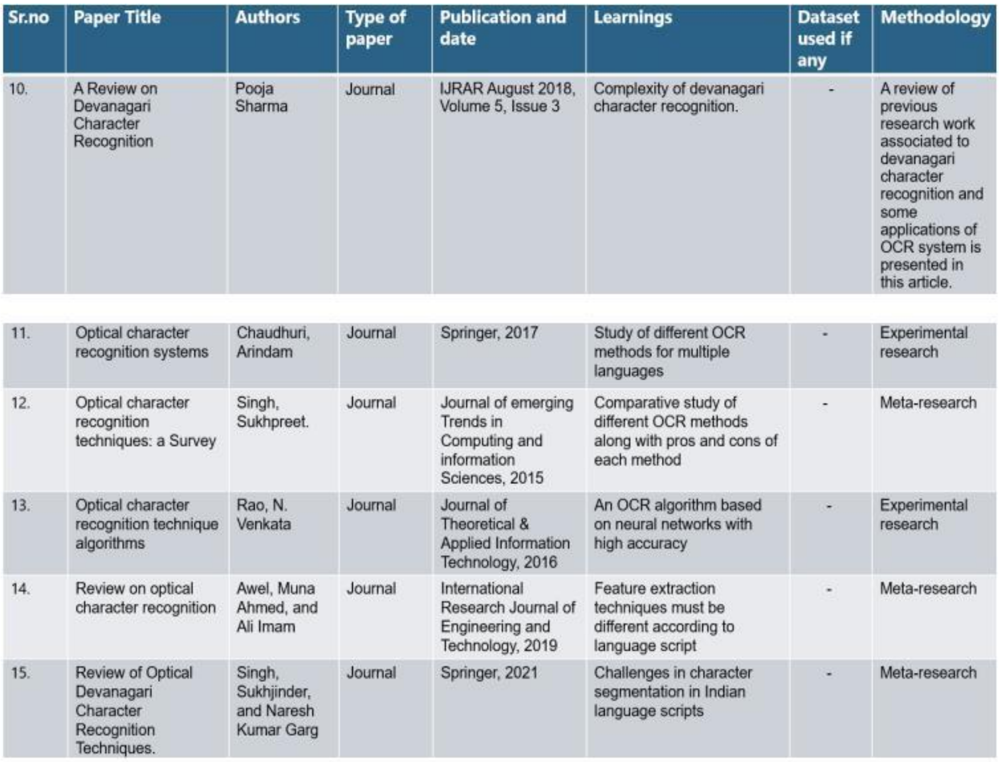
Those languages are Hindi, Sanskrit, Kashmiri, Marathi, and many more. English character recognition is mostly studied by researchers and a lot of commercial systems are used for it. But for Indian languages, the research work is very limited because of the complex formation of the language.Optical Character Recognition (OCR) is one such capability, which involves the automated conversion of scanned documents/images of machine-printed/handwritten characters into machine-readable digital form. OCR systems aid in the automated processing of large amounts of textual data.

II. LITERATURE SURVEY



Sr.no.	Paper Title	Authors	Type of paper	Publication and date	Learnings	Dataset used if any	Methodology
1.	The State of the Art in On-Line Handwriting Recognition	Tappert, Charles C., et al.	Journal paper	IEEE - 2015	Handwriting properties, External Segmentation, Pre and post processing	Self generated data set	External segmentation
2.	Handwritten character recognition through two-stage foreground sub-sampling	Vamvakas, Georgios, et.al	Journal paper	Elsevier - 2016	SVM Classifier	CEDAR Character Database, MNIST	Recursive subdivisions of the character image
3	Diagonal based feature extraction for handwritten character recognition system using neural network	Pradeep, J and Srinivasan, E and Himavathi, S	Conference Paper	IEEE (3rd international conference on electronics computer technology) 2015	Feed forward neural networks	Unknown	Diagonal feature extraction
4.	Deep learning for trilingual character recognition	Yashodha, M and Niranjan, SK and Aradhya, VN Manjunath	Journal paper	IGI (International Journal of Natural Computing Research) - 2019	Weight Regularization, Sparsity Regularization & Sparsity Proportion	Public dataset by HP labs	Auto Encoders Based Model
5.	A new method for line segmentation of handwritten Hindi text	Garg, Naresh Kumar and Kaur, Lakhwinder and Jindal, Manish Kumar	Conference paper	IEEE (seventh international conference on information technology) 2017	Line Segmentation	-	-
6.	A Complete Optical Character Recognition Methodology for Historical Documents	Vamvakas, Georgios, Basilis Gatos, and Stavros J. Perantonis, Nikolaos Stamatopoulos	Conference Paper	DBLP Computer Science Bibliography, IEEE, September 2018	Use of clustering for making an OCR	-	A clustering scheme is adopted in order to group characters of similar shape.

Fig.2.1 OCR Implementations



Sr.no	Paper Title	Authors	Type of paper	Publication and date	Learnings	Dataset used if any	Methodology
10.	A Review on Devanagari Character Recognition	Pooja Sharma	Journal	IJRAR August 2018, Volume 5, Issue 3	Complexity of devanagari character recognition.	-	A review of previous research work associated to devanagari character recognition and some applications of OCR system is presented in this article.
11.	Optical character recognition systems	Chaudhuri, Arindam	Journal	Springer, 2017	Study of different OCR methods for multiple languages	-	Experimental research
12.	Optical character recognition techniques: a Survey	Singh, Sukhpreet.	Journal	Journal of emerging Trends in Computing and information Sciences, 2015	Comparative study of different OCR methods along with pros and cons of each method	-	Meta-research
13.	Optical character recognition technique algorithms	Rao, N. Venkata	Journal	Journal of Theoretical & Applied Information Technology, 2016	An OCR algorithm based on neural networks with high accuracy	-	Experimental research
14.	Review on optical character recognition	Awel, Muna Ahmed, and Ali Imam	Journal	International Research Journal of Engineering and Technology, 2019	Feature extraction techniques must be different according to language script	-	Meta-research
15.	Review of Optical Devanagari Character Recognition Techniques.	Singh, Sukhjinder, and Naresh Kumar Garg	Journal	Springer, 2021	Challenges in character segmentation in Indian language scripts	-	Meta-research

Table 2.1: Literature Survey

DOI: 10.17148/IJARCCE.2023.12402

III. PROPOSED SYSTEM

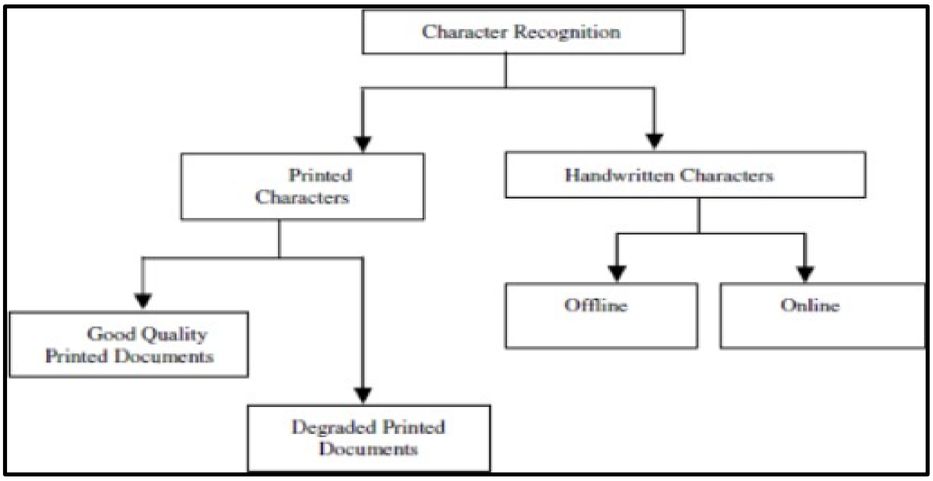


Fig .3.1 OCR implementations

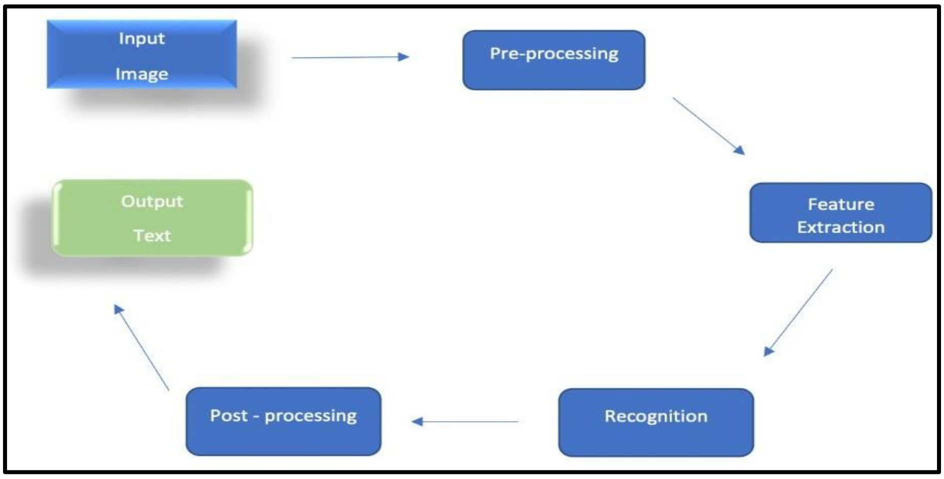


Fig.3.2 proposed solution

1. Application should take the input in image format containing some text in Devanagariscript.
2. Application should be able to convert the given input image into editable text.
3. Application should output the editable text which can then be used by the user.
4. Application should also be able to identify and convert handwritten text document

IV. PROBLEM STATEMENT

To design and implement a recognition system for handwritten Devanagari Script using supervised learning

V. GOALS AND OBJECTIVES

1. Design and implement OCR for Devanagari characters
2. Segmentation of Devanagari Script into sentences, words and characters.
3. Implement OCR for paragraphs
4. Deploy the OCR model
5. Create an application with the deployed model

DOI: 10.17148/IJARCCE.2023.12402

VI. SCOPE AND MAJOR CONSTRAINTS

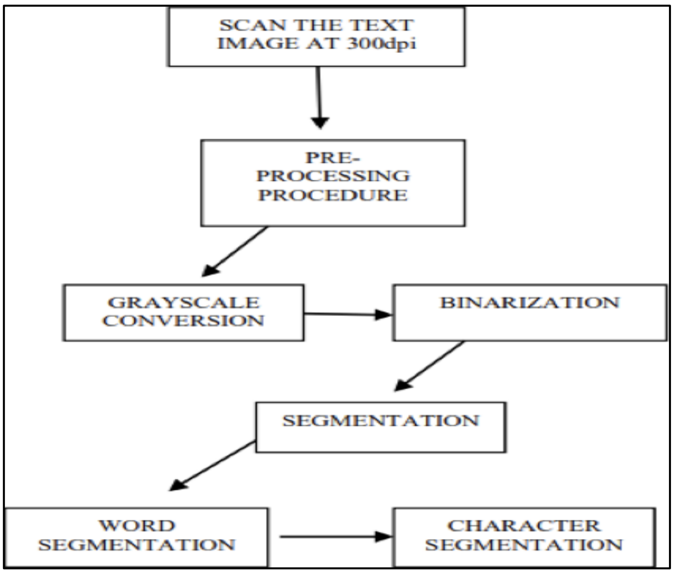


Fig.6.1 scope of project

scope:

1. Pre-processing and segmentation is the part in which the existing models lack behindand needs to be worked upon.
2. Model to be more focused on the complexities of Devanagari script.
3. Increase in the overall efficiency of the model.
4. Deployment of a simpler Web application with our model with image to documenttranslation.

Constraints:

The complexities of the Devanagari Script such as connected letters, upper zone, lower zone etc, makes the feature extraction a bit tedious. Finding appropriate datasets for the same, is a difficult constraint to overcome.
After the feature extraction, converting them into raw text will be another constraint.

VII. SYSTEM DESIGN

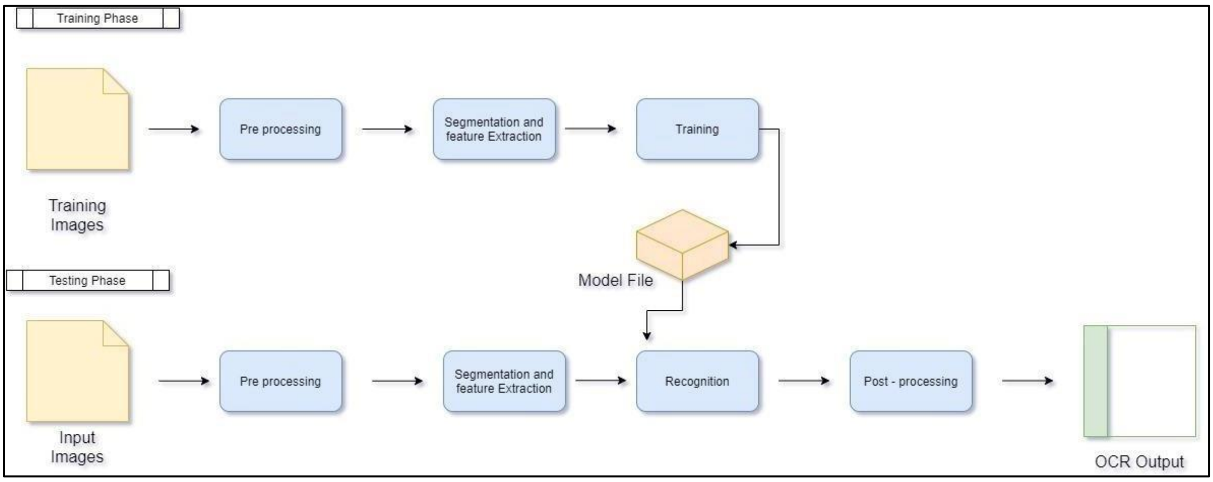


Fig.7.1 system architecture

VIII. ARCHITECTURE OF SYSTEMS

A. Segmentation and Classification

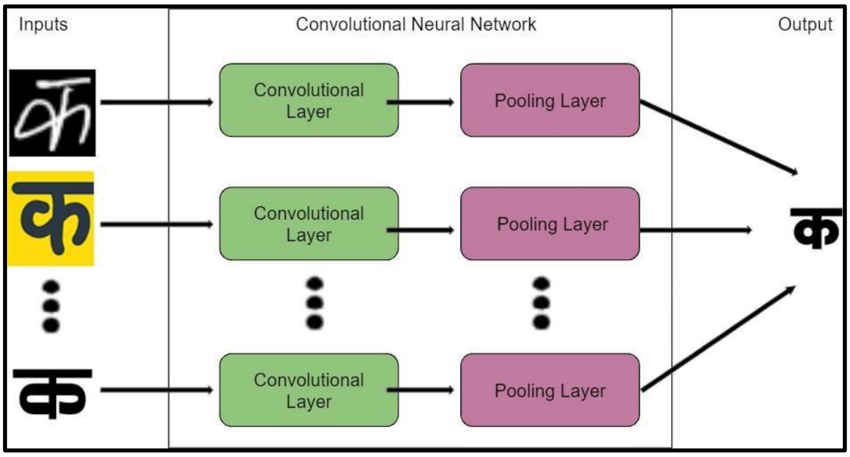


Fig.8.1 CNN for character recognition

As discussed earlier, segmentation is especially challenging in the case of Devanagari script ascompared to English or any other language script because of its characteristics. In such scripts,a text word may be partitioned into three zones. The upper zone denotes the portion above the headline; the middle zone covers the portion of basic and compound characters below the headline and the lower zone that may contain some vowel and consonant modifiers. Theimaginary line separating the middle and lower zone may be called the baseline.

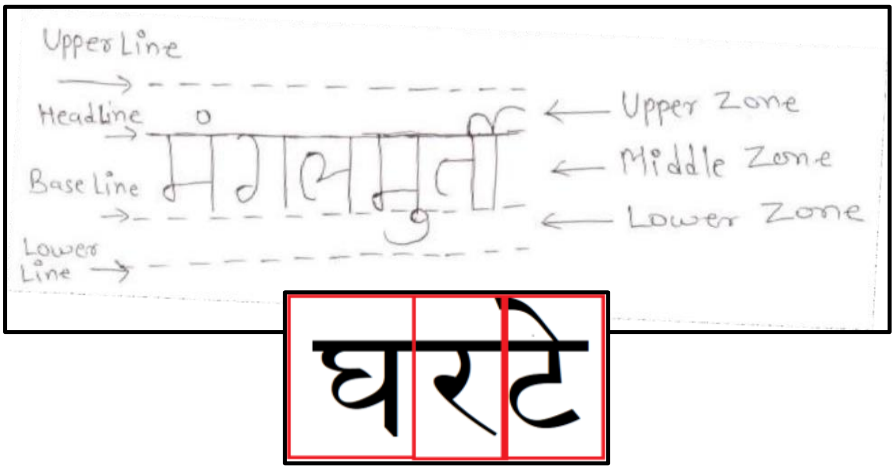


Fig.8.2. Complexities of Devanagari Script

Segmentation is done by dividing a given word into different partitions as well as segregatingthe individual letters in those portions. After all the individual letters and modifiers are identified, they are fed to a classification algorithm which will output the Unicode identifier for that particular character. For better classification results a sequential CNN algorithm is preferred.

DOI: 10.17148/IJARCCE.2023.12402

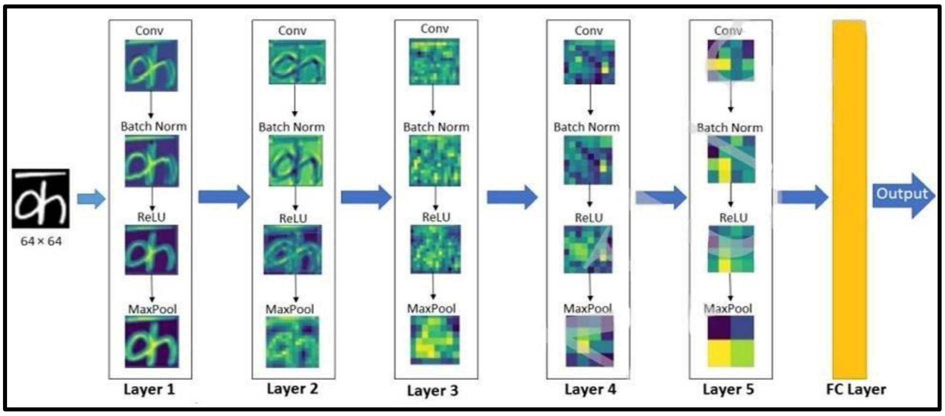


Fig.8.3.CNNN layers

A series of convolution operations along with/without pooling and non-linearity activation function are applied on the given input. The filters are applied in the CL to extract relevant features from the input image to pass further. Each filter gives a different feature for correctprediction. To retain the size of the image, it helps reduce the number of features taken froman image. The convoluted output is obtained as an activation map.

IX. FUTURE SCOPE AND CONCLUSION

A lot of different techniques have been published by authors in this field but still there are noproper effective OCR systems in application right now. A lot of it comes down to thecomplicated nature of Devanagari script in general. After going through multiple research papers, we narrowed down our search to two different techniques. After evaluating them with regards to feasibility, effectiveness and simplicity. We decided to go with the second approach as it is much more flexible and gives better overall results as compared to the first according to the literature in this field. This technique is based on an encoder decoder modeland uses CNN, RNN, and LSTM for performing character recognition. This technique gave us an accuracy of about 88% and is deployed using Flask. Other factors affecting the accuracy such as brightness, contrast, shadows etc. can be worked upon, to increase the accuracy further. Better Image pre-processing can be done such as reducing background noise to handle real time images more accurately. This way a decently accurate OCR system can be built for handwritten Devanagari script recognition.

REFERENCES

[1]. Tappert, Charles C., Ching Y. Suen, and Toru Wakahara. "The state of the art in online handwritingrecognition." IEEE Transactions on pattern analysis and machine intelligence 12.8 787-808.

[2] Vamvakas, Georgios, Basilis Gatos, and Stavros J. Perantonis. "Handwritten character recognition through two-stage foreground sub-sampling." Pattern Recognition 43.8 2807-2816.

[3] Pradeep, J., E. Srinivasan, and S. Himavathi. "Diagonal based feature extraction for handwritten character recognition system using neural network." 2011 3rd international conference on electronics computer technology. Vol. 4. IEEE.

[4] Vamvakas, Georgios, et al. "A complete optical character recognition methodology for historical documents." 2008 The Eighth IAPR International Workshop on Document Analysis Systems. IEEE.

[5] Karthick, K., et al. "Steps involved in text recognition and recent research in OCR; a study." International Journal of Recent Technology and Engineering 8.1 (2019): 2277-3878.

[6] Wang, Zilong, et al. "LayoutReader: Pre-training of Text and Layout for Reading Order Detection." arXiv preprint arXiv:2108.11591 (2021).

[7] Yashodha, M., S. K. Niranjan, and VN Manjunath Aradhya. "Deep learning for trilingual character recognition." International Journal of Natural Computing Research (IJNCR) 8.1 (2019): 52-58.

[8] Garg, Naresh Kumar, Lakhwinder Kaur, and Manish Kumar Jindal. "A new method for line segmentation of handwritten Hindi text." 2010 seventh international conference on information technology: newgenerations. IEEE, 2010.

[9] Thakur, Anupama, and Amrit Kaur. "Devanagari handwritten character recognition using neuralnetwork." Int. J. Sci. Technol. Res 8.10 (2019).

[10] Chaudhuri, Arindam, et al. "Optical character recognition systems." Optical Character Recognition Systems for Different Languages with Soft Computing. Springer, Cham, 2017. 9-41.