IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EXPLORING AND COMPARING METHODS FOR HANDWRITTEN DEVANAGARI CHARACTER IDENTIFICATION

1] Shiddesh Thube **2] Mr Peeyush Pareek**
3] Dr. Arshiya Khan **4] Dr. Prof. R Deshpande**
MCA Faculty of Science and Technology, JSPM University

ABSTRACT

it is not easy to read textual information in scanned images when the font style or the writing script is different a lot of work has been done on English text though Indian scripts have not been given much attention one of the most common is Devanagari used in Hindi Sanskrit Marathi and Kashmiri in the recent past scholars have been engaged in attempting to decode Devanagari letters the contribution of the present study is that it compares the four data-driven models namely Multi-layer Perceptron k-nearest neighbors support vector machine and random forest with two feature selection methods deep Convolutional model and histogram of oriented gradients the comparison will help in future research in getting better accuracy in recognition of scanned text in Indian languages

***Keywords*- Structural recognition, Image-based text extraction, Method comparison, Artificial neural network, KNN classifier, SVM algorithm, Image gradient histogram.**

I. INTRODUCTION

advanced technology development combined with powerful computing systems allows computers to enhance their intelligence for executing tasks which traditionally required human labor multiple researchers have devoted years to developing improved methods for computer-human interaction thus producing various helpful tools and technologies the important tool within this field is vision-based pattern recognition of text because it enables conversion of message from scanned documents or images to editable machine-understandable files ocr stands as an essential innvation because it enables the computerized handling and processing of substantial text data volumes the majority of ocr technologies function with one particular script whereas multiple scripts within a single document remain difficult to identify the development of an effective ocr system represents a complex yet essential task for india because it needs to deal with 23 official languages and 13 different scripts in a multilingual setting document recognition spanning different indian scripts proves difficult because present-day ocr programs handle only single scripts the insufficient funding and market interest in Ocr research has not hampered the ongoing efforts to advance this field because document electronic processing needs continue to expand the research of Ocr technology toward the Devanagari script which serves several Indian languages including Hindi alongside Marathi and Sanskrit and Nepali and others commenced during 1977 the recognition of handwritten Devanagari characters has gained more attention during recent times because this script continues to be widely used despite its diverse writing styles the research provides essential information about different methods that scientists developed during multiple years to detect handwritten Devanagari characters the paper reviews multiple research approaches while examining their influence on enhancing Devanagari Ocr performance efficiency along with Accuracy outcomes

II. CHARACTER RECOGNITION

character identification is an important area of research in the field of computer science for many years it mainly deals with the process of converting printed or handwritten letters from scanned papers or images into an editable and readable digital format using special software or tools this process is commonly known as optical text recognition otr where scanned images of characters are detected and changed into text that computers can easily understand devanagari is most popular and widely used scripts in india it is used mostly for writing many indian languages like hindi sanskrit marathi and kashmiri millions of people use this script in their daily communication and writing character recognition is generally classified into two main categories handwritten text recognition and computer-typed text recognition further computer typed text can also be divided based on the quality of the text like high-quality text clear print and low-quality text blurred or unclear print as shown in figure 1 handwritten character recognition is divided into two types offline and online offline recognition means recognizing the writing from scanned images or documents online recognition means detecting characters directly while writing using digital tools like touch screens or stylus pens in this paper we mainly focus on the offline recognition of devanagari 4 5 script characters in the devanagari writing system the text can be divided into three zone the top zone middle zone and bottom zone the top zone lies above the headline known as shirorekha the middle zone contains the main part of the character and the bottom zone is the area below the baseline which sometimes includes extra marks or symbols segmentation is the process in which words are broken down into smaller parts so that each letter and symbol can be separated clearly after separating all the characters and symbols they are sent for classification where a specific code unicode is given to every recognized character finally the output is produced in the form of readable and editable digital text

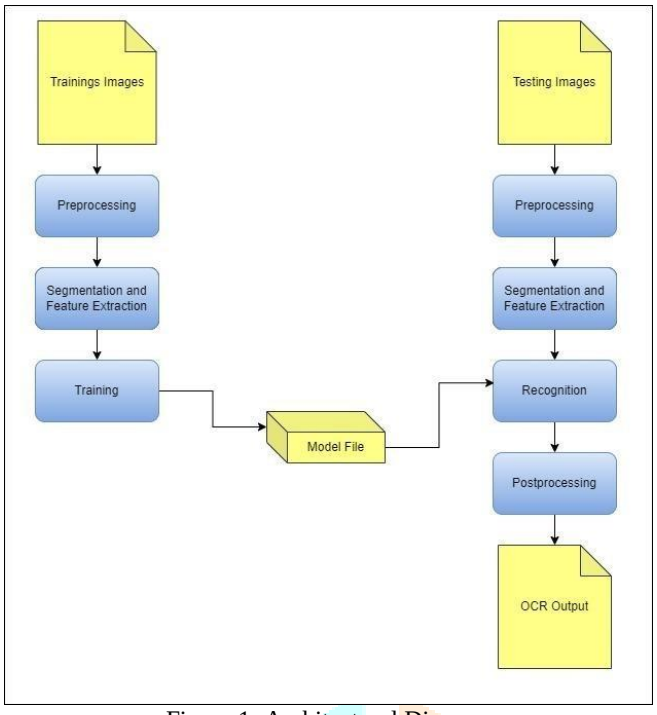


Figure 1: Architectural Diagram

III. DEVANAGARI SCRIPT

it is primarily utilized for writing several indian languages utilized for composing various indian languages including hindi indic script of sanskrit marathi and nepali differs distinctly from english-based composing systems it does not use uppercase or lowercase letters all characters are formed in same size and formhindi utilizing this indic script ranks as the third highest number Linguistic population globally following chinese and english includes 13 vowels and 34 obstruent sounds sometimes two or more obstruents are combined to form a new character shape these special combinations are known as compound characters in addition various marks or figures may be added to alter the sound of symbol text written in devanagari can be devided into three regions the upper region the middle region and the lower region the upper zone is the area above the main writing line in which some signs are written most of the characters are in the middle zone and the lower zone is situated below the main line and it can have extra marks or symbols the horizontal line between the middle and lower areas is termed as the baseline segmentation involves dividing a word into small units or separating the characters explicitly this is necessary in providing computers or machines with the ability to identify each letter precisely after every symbol has been determined a special number referred to as unicode is given to it to enable the text to appear appropriately on digital devices including computers and mobile phones represents the boundary between the middle and lower areas is referred to as the baseline segmentation is the act of dividing a word into small units or separating each letter distinctly this action is necessary because it makes the computers or machines to recognize each letter easily after distinguishing each letter and symbol a special code called unicode is assigned to each character this code aids in the proper display of the text on computer and mobile phones

IV. LITERATURE STUDY

many researchers have done studies on devanagari ocr optical character recognition and their important findings are explained here in a simple way one study focused on how to recognize the shapes of letters it explained different methods used before and after the recognition process it also said that online learning is very helpful because it allows the system to adjust according to different users another method was suggested for recognizing handwritten letters using a technique that breaks the image into smaller parts for better feature extraction another research discussed using a neural network a type of artificial intelligence for offline character recognition it showed that using diagonal feature extraction gives better results than using only horizontal or vertical features one more approach was designed to work on both machine-printed and handwritten documents it didnt require any fixed font style or special database making it easy to use for different types of documents some research explained that it is possible to get faster and better results with less processing time it also showed that recognizing letters from multiple languages can be done more effectively in another study a large dataset of 500000 document images was created this dataset helps to check the reading order of text in documents they also introduced a layout reader tool that helps in identifying the correct reading order in ocr systems further research presented a deep neural network method for recognizing characters from devanagari kannada and english scripts one more method was based on identifying header lines baselines and the shape of the letters this approach was useful in separating lines properly even if the handwritten text was not straight another study talked about the difficulty of recognizing Devanagari letters because of extra signs and symbols to handle this researchers suggested breaking words letter-by-letter instead of dealing with all modifiers together lastly some Important Pre-processing steps like removing noise detecting skew thinning the text and segmenting the characters were introduced one method called fmrf showed the best accuracy in different situations like loop intersections and overlapping lines

V. FEATURE EXTRACTION

Various solutions aiming to recognize handwritten Devanagari script letters have been developed throughout the recent years. Segmentation serves as the main approach for carrying out this operation. Words need to be segmented through the process of dividing them into individual parts containing separate letters for proper segmentation. English text shows favorable segmentation because its single characters divide expediently. Devanagari script shows most of its letters as connected units creating obstacles during the separation process.

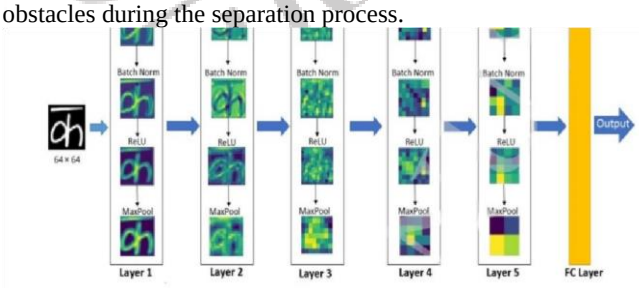


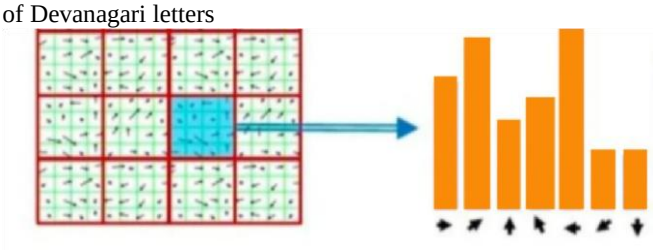
Figure 4: CNN layers

figure 4 layers of Cnn special features of individual letters will be identified after the process of letter segmentation these particular features are all that is needed in the system framework to thrive in letter identification the operation can be carried out in a number of user procedures the aim demands an array of famous means to achieve the success 1convolutional neural network features one of the models is referred to as Cnn which operates on features that emulate human intelligence as well as visual processing abilities the tool uses its visual abilities to process the information in the pictures that are then articulated into a visual comprehension there are several confirmation processes that Cnn image processing system undergoes the system identifies various patterns which include straight lines as well as curved edges and sides which are present in image data the operation of image scanning in this technique employs four layers with two extra layers of preserving details as well as rescaling the image the system has an improved efficiency in

information processing in the recognition and classification of letters

B) Histogram Oriented Gradients

The second implementation of shape recognition occurs through HOG. The detection process of the system depends on color intensity variations which usually occur in places of letters placement the changes allow one to understand letter forms based on the generated necessary data the program has an ability of identifying suitable letters without being influenced by the variation of writing style the computer systems use the ensemble of hog and Cnn to deal with hand written and printed recognition of Devanagari letters



VI. CHARACTER CLASSIFICATION

Lettertagging with various algorithms machine learning is applied in this project was applied to classify various types of letters a number of algorithms were applied to evaluate the speed and accuracy with which they could process data and give the outcome various features were extracted with the help of these algorithms 324 features were extracted on images based on training time and accuracy 64 features were generated on images based on histogram of edge directions orientated directional changes although hog mlp and random forest were faster in processing data cnn information extraction took longer time to train the features used in the isolation and the algorithm implemented showed accuracy of results but certain errors were identified too because of the drawbacks of the chosen algorithms and features

IX. CONCLUSION

The results are clearly demonstrating that image-processing neural networks are far better in image grouping especially when a large quantity of labeled training data is available an accomplishment that would not be possible without the diligent efforts of researchers who have prepared and made available such Datasets in the scenario where large Datasets are scarce especially in case of some languages the histogram of oriented gradients hog feature is appearing as a good alternative to character grouping support vector machines Svm have already proven to be better in comparison to other models over the years Devanagari character recognition has indeed progressed well modern algorithms can now recognize characters with a great

X. REFERENCES

[1]. Tappert, Charles C., Ching Y. Suen, and Toru Wakahara. "The state of the art in online handwriting recognition." 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. "Diagonalbased 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 opticalcharacter 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 recognitionand 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 ofText and Layout for Reading Order Detection." arXiv preprint arXiv:2108.11591 (2021).

[7] Yashodha, M., S. K. Niranjan, and VN ManjunathAradhya. "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 ManishKumar Jindal. "A new method for line segmentation of handwritten Hindi text." 2010 seventh international conference on information technology: new generations. IEEE, 2010.

[9] Thakur, Anupama, and Amrit Kaur. "Devanagarihandwritten character recognition using neural network." Int. J. Sci. Technol. Res 8.10 (2019).

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

[11] Singh, Sukhpreet. "Optical character recognitiontechniques: a survey." Journal of Emerging Trends in Computing and Information Sciences 4.6 (2013): 545550.

[12] Rao, N. Venkata, et al. "OPTICAL CHARACTER RECOGNITION TECHNIQUE ALGORITHMS." Journal of Theoretical & Applied Information Technology 83.2 (2016).

[13] Awel, Muna Ahmed, and Ali Imam Abidi. "Review on optical character recognition." International Research Journal of Engineering and Technology (IRJET) 6.6 (2019): 3666-3669.

[14] Satav, M. S., Varade, T., Kothavale, D., Thombare,S., & Lokhande, P. (2020, November). Data Extraction From Invoices Using Computer Vision. In 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS) (pp. 316-320). IEEE.

[15] Chikmurge, D., & Shriram, R. (2019, December). Marathi handwritten character recognition using SVM and KNN classifier. In International Conference on Hybrid Intelligent Systems (pp. 319-327). Springer, Cham.

[16]S. Acharya, A.K. Pant and P.K. Gyawali â€œDeepLearning-Based Large Scale Handwritten Devanagari Character Recognition, In Proceedings of the 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pp. 121-126, 2015.