ISSN 2395-1621

Devanagiri Text Digitisation Using Transfer Learning on Handwritten Text

Ankit Das, Pranav Adkine, Yashashree Abnave, Rohan Purkaith

ad53ankitdas@gmail.com
adkinepranav@gmail.com
abnaveyashashree@gmail.com
purkaithrohan@gmail.com

Department of Computer Engineering
D.Y Patil Institute of Technology, Pimpri, Pune-18.

ABSTRACT

With the advance of computing and technology, manual systems are getting replaced. The purpose of the system is to automate the existing manual system with the help of a web application so that valuable data can be stored for a longer period with easy accessing and manipulation of the same. This system can lead to error free, secure, reliable and fast management system. Western languages have powerful recognition engines powered by google, but devanagiri that is popular script in India, does not have proper digitisation tools. Here we look into the available techniques that have already been applied and improve the rate of recognition ting CNN transfer learning. DHCD (Devanagari Handwritten Character Dataset) is an open dataset available for the purpose, it has 46 classes cache with 2000 samples in each class. The dataset needs additional classes to cover all the possibilities of characters, those are intended to be added and a wholesome dataset will be made. The unknown words or conflicts can be handled by the user

Keywords— Devanagari Image to Text ; Transfer Learning; Devanagiri Text Digitisation ; CNN for Devanagiri Text Recognition ; Alexnet.

ARTICLE INFO

Article History

Received: 10th March 2020

Received in revised form :

10th March 2020

Accepted: 13th March 2020

Published online :

13th March 2020

I. INTRODUCTION

There is a large requirement of a technology that converts a handwritten texts in digital text format, mostly used for digitising old handwritten documents in NIC (National Informatics Centre), Courtrooms, Registrar offices etc. Offices like NIC faces issues like storage of large volume of data, documents takes a lot of physical space, which tends to be expensive and limited. Searching and indexing is required to find a data, which itself is another set of task. With time, there are high chances of paper quality getting deteriorated, ink quality also gets deteriorated, making it difficult to read and understand the information. As the storage of document is in physical spaces miss happenings like fire, flood, earthquake or other natural calamities can occur. Manual manipulation becomes easy for the physical data; evidence tampering can be done very easily. There are technologies discovered for digitising a printed text in an ideal format, but there are not many software in the market to digitise a handwritten document with good accuracy. If the digitisation of devanagiri text recognition becomes successful, it will be very useful in storing the data for even a longer period of time, around 10,000 files which requires a large room for storage will take around 1 TB in the digital space. It will be useful in transferring the data from one place to another; the data can be randomly accessed from any location with the proper authorisation. The digitisation of data also helps in reducing the chances of evidence tampering as the digital medium keeps a full-proof record of the transactions occurring. Important documents can be distributed and stored at multiple places, so that it becomes difficult for an unauthorised user to access the data. The research work possesses a digitised document of the handwritten devanagiri script. Using transfer learning it will become easier to train with more number of classes for compound characters used in devanagiri. It provides a faster conversion and takes less time to train. The features like conflict resolution is provided for yet less ambiguity. The proposed interface is a web app to implement the frontend, it can also be implemented as an API.

II. LITERATURE REVIEW

[1] Segmentation is the process of dividing the handwritten text words into various segments. In the paper suggested by Shafali Goyal and Akash Bhatla handwritten words are segmented into upper zone, middle zone and lower zone to recognize various characteristics of word. The text document is scanned and pre-processed by setting the threshold value of image to 200, converting it into binarized

form i.e in the form of 0's and 1's and removing the noise from the images. The image document is divided into vertical stripes each of 100 pixels and each vertical stripe into horizontal direction. The gaps between two corresponding lines are calculated by checking the pixels of each line in horizontal direction. The line is ignored if there are black pixels but if the whole pixels are white then it is considered as space or gaps and stored in array to find the mid-points of these gaps. If the difference between two mid-point is more than the expected value then the average of these mid-points calculated. The line is ignored if it contains black pixels. After combining the segmentation result lines are drawn at the calculated mid-point to represent segmentation of lines and performance analysis of the result is performed. The accuracy of this method was found to be 95% after testing on various digital documents.

[2]Projection profile based algorithm is described in the paper by Rahul Garg and NareshGarg where projections of pixels are used to segment the text lines. They have used piecewise projection profile to divide the whole document into 6, 7 & 8 strips and have calculated result based on many different document images to deal efficiently with skewed text and overlapped and touched lines. The algorithm binaries the image, calculates the height and width of the image and divides the image into vertical strips and calculates the number of black pixels in each row. This algorithm calculates the gap by assigning colours to the pixels and dividing them accordingly. Overlapping and touching of character cause disruption in the algorithm. The accuracy of this algorithm is 93.2%.

[3]In the method proposed by RenuDhir header line and base line is calculated for the line segmentation. The header line has maximum number of black pixels, whereas base line has aminimum number of pixels in a row. Before finding the header lines and base lines , the average line height of is found and given to the algorithm which finds baseline and header line as an input. The rows are divided into equal half and each half is used to find the header lines and base lines.

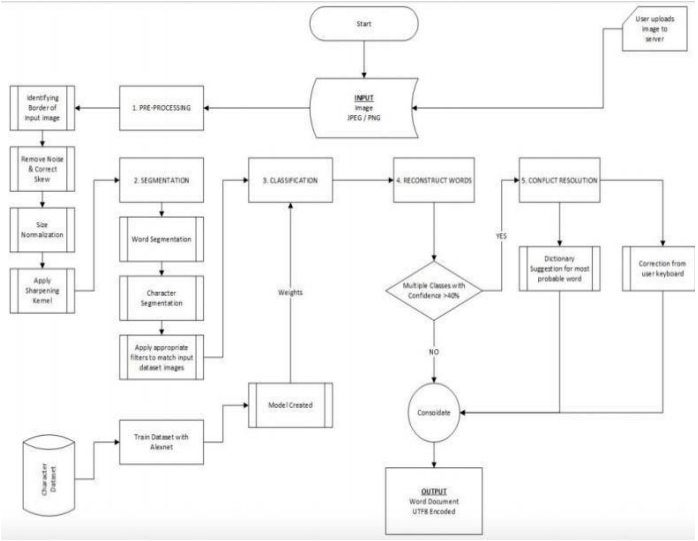
[4]In the paper by Lajish VL and Sunil Kopparapu character recognition is based on a set of primitives hand written strokes that can be used to write the complete alphabet set in Devanagari. This paper proposes to use extended directional feature(EDF) set for the recognition of primitives. The motivation for using strokes is influenced from the speech recognition literature. Strokes are used to plot the points on a letter. The number of points varies depending on the size of stroke and also the speed with which it was written. Digitization device sample the data uniformly per time and smoothen the points. Discrete Wavelet Transform is used to identify the curvature points from the smoothed handwritten data. These curvature points are used to identify the character. The average accuracy of this method is 65.6 percent.

[5] In the paper By AkhilDeshmukh, Rahul Meshram, Sachin Kendre, Kunal Shah Character recognition is divided into three steps: Image correction, Segmentation and Recognition. a. 1) Correction: Principle component analysis (PCA) is used to discover the hidden part and to obtain the clear information of the complex data in the image. It consists of Noise reduction, Gray scaling, Binarization. 2) Segmentation: Line segmentation, Character Segmentation, Word Segmentation and Fused WordSegmentation is used for segmentation process and character recognition. 3) Recognition: For recognition Eigen space algorithm is used using the concept id Gerschgorin's theorem which is usually used for face recognition. For the comparisons 5 different stored samples are used.

[6] This paper by ShaileshAcharya,Ashok Kumar Pant and Prashnna Kumar Gyawali proposes to use Deep Learning architecture to distinguish the characters in DHCD dataset. Devanagari Handwritten Character Dataset is set of 92 thousand images of 46 Devangari characters. Deep Neural Network works on raw pixels data generating the best features and using them to classify the inputs of different classes. Convolutional Neural Network a class of Deep Neural Network works with a small set of parameters and is easier to train and has ability to correctly model the input dataset by changing the number of hidden layers and the trainable parameter in each layer and also make the correct assumption based on the nature of images.

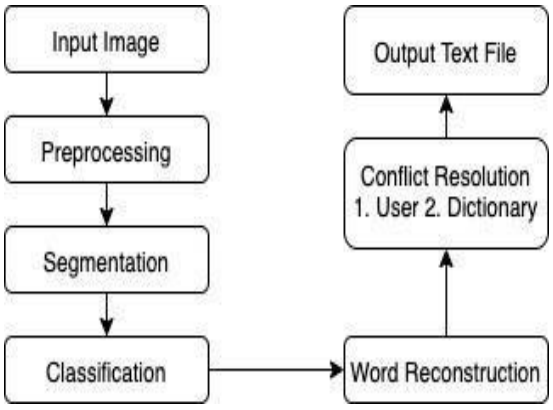
[7] In the method proposed by Prasad.K.Sonawae and ShushmaShekle , handwriting devnagiri text is classified using transfer learning mechanism with the help of Alexnet. Alexnet is a convolutional neural network. It is trained over a dataset of 16879 samples of 22 consonants of devanagari script. After performing preprocessing and extraction of feature from the data , classifier is trained . The classifier is used to classify the data into a number of classes. The accuracy of these methods depends on the selected features for training. For training the network dataset is divided into 3 parts: training, validation and testing. This method uses MATLAB2017 and neural network toolbox for implementation. Using transfer learning 94.49% validation accuracy and 95.46% test accuracy was found.

III. PROPOSED SYSTEM



The major components of our proposed system are

1. Pre-Processing
2. Segmentation
3. Classification
4. Word Reconstruction
5. Conflict Resolution



The image of the document will be taken as input, pre processing will be done, it includes identifying borders, removing noise and correcting skew, size normalisation and applying sharpening kernel. The next major step after the image is processed is segmentation of the image into the words, we can use Tessaract-ocr which is open source, after that we will get a stream of characters that will be passed on to the next stage. The classification stage will classify the characters into different devanagiri characters that are available in the Dataset used for training i.e. Devanagiri Handwritten character dataset. The prediction of word will be based on reference from the dictionary. Once the character is predicted to be something, it will get some confidence score. Instead of having the best prediction as the output we will take the top 2 as final, but if a single class has confidence over 60% we will take it as final. After all the classes are found for a single word, all the possibility will be checked in dictionary. The best fit will be chosen as the final word. Now, if there is any ambiguity in the word a final edit option will be provided to the user to make changes as per his/her convenience.

IV. CONCLUSION & FUTURE WORK

A conclusion: The system scans the document Handwritten in devanagiri script.It converts the document into digitized form with an improved accuracy.

Future Work: Adding conflict resolution ,i.e a feature that helps you to edit the documents after it is converted into a digitized format.

REFRENCES

[1] Segmentation - Shafali Goyal,Akash Bhatla

[2] Garg, R. and Garg, N.K., 2014. An algorithm for text line segmentation in handwritten skewed and overlapped Devanagari script. International Journal of Emerging Trends in Engineering and Development, 4(5), pp.114-118

[3] Jangid, M. and Srivastava, S., 2018. Handwritten devanagari character recognition using layer-wise training of deep convolutional neural networks and adaptive gradient methods. Journal of Imaging, 4(2),

[4] Lajish, V.L. and Kopparapu, S.K., 2014, December. Online handwritten Devanagari stroke recognition using extended directional features. In 2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS) (pp. 1-5). IEEE

