NATIONAL INSTITUTE OF TECHNOLOGY, SILCHAR

Summer Internship June-July 2022

Under

Dr. Ujwala Baruah

Dept. of Computer Science and Engineering

By

1. Biplob Phukan (1912129)
2. Rohit Kumar (1911049)
3. Priyanshu Maurya(1911051)
4. Satyabrat Borah(1911073)

RESEARCH PAPER

Offline Odia handwritten character recognition with a focus on compound character

Authors: Raghunath Dey, Rakesh Chandra Balabantaray, Sanghamitra Mohanty

Date: 3 January 2022

ABSTRACT

There are many languages which aren’t yet worked on and thus lacking modern touch to them. Recognition of such languages’ e,g. Odia, Assamese etc. character images using offline OCR was the field of research. The developed algorithm claims to achieve a recognition accuracy of 86.56% on the dataset with 112 classes of characters (where each class means individual character). Our implementation of the proposed algorithm yielded us accuracy of 70.93% on the IIITB Odia Char dataset.

INTRODUCTION

A machine can recognize handwritten text using optical character recognition (OCR).

OCR (optical character recognition) is the use of technology to distinguish printed or handwritten text characters inside digital images of physical documents, such as a scanned paper document. The basic process of OCR involves examining the text of a document and translating the characters into code that can be used for data processing. OCR is sometimes also referred to as text recognition. There are two ways to recognize handwritten texts: online and offline. In the case of online, some electronic devices can be used to trace the writing direction when the writing is going on, such as online signature authentication. On the other hand, offline handwriting identification performs the identification of text from scanned textual images. It includes the recognition of numbers from bank cheques, addresses from letters, etc. With the help of this, it is possible to save both space and time. Here, there is no information available about the movement of the pen tip, the trajectory, or the direction of the text line. Offline OCR is, therefore, more critical than online OCR.

Several scripts are used in India for various languages. There is considerably less amount of research towards OCR present in the literature to support handwritten text identification. Furthermore, it can be noted that most of the studies attempt to recognize Devanagari and Bangla handwritten characters in comparison to other regional languages.

Saving the contents as a text file form is a better choice than a scanned copy of the pages to avoid running out of storage space. The manual conversion of this job is impractical. It’s essential to automate the conversion of these image files to text files, which necessitates the use of offline OCR for Odia characters. The lack of Odia databases to train the OCR engine is the primary impediment. The performance of Odia OCR must be improved to fulfil the requirements of real-time recognition. It motivates strongly to donate a good quality of Odia characters and numeral dataset to the research community working in this domain. Handwritten character recognition is among the several popular applications of computer vision. Though, this job is not simple. Researchers are encouraged to enhance recognition results using various pattern recognition algorithms, which helped the offline text retrieval and digitization process.

Smartphones in India are expanding exponentially in all the regions of India, which will lead to heavy man-computer interactions in the future. OCR developments for Indian scripts, including popular scripts such as Odia, will be in high demand. As a result of these factors, a high-quality Odia handwritten text with various modifiers and compound character recognizers is now required. It necessitates the collection of a standard handwritten Odia database consisting of different possible compound characters. The current study proposes a character database as well as a hybrid feature extraction technique. The neural networks models are designed to take extracted features as input instead of using direct images. It would speed up the character recognition task because it would take less time to process the extracted features than compared to immediate images. Thus, we are trying to implement this process to obtain a digital copy of the old documents and manuscripts, which will be very useful in real life.

LITERATURE SURVEY

The research on handwritten character recognition began lately on some unexplored languages like Odia (a Dravidian language). For Odia only, there are around 35 papers concerning offline recognition of handwritten characters.

Tripathy et al. 2003 suggested a method to recognize the numerals, and the authors have used threshold-based binarization as a primary preprocessing technique. Features are extracted based on reservoir area and location, the path of water flow, the number of loops, the center of gravity, ratio between reservoir and loop, profile-based features, on jump discontinuity. Finally, using a binary tree classification approach, the accuracy obtained was around 97.74%.

The paper Roy et al., 2005 suggested computing region of interest (ROI) based on bounding box and segmented the characters into blocks. The feature extraction was based on a chain code histogram having a dimension of 400. By applying neural networks and quadratic classifiers, an accuracy of 94.81% was reported in the study. Except for these preprocessing techniques, few others like Gaussian filter, Roberts filter are applied in and accuracy of 98.54% was achieved using the same quadratic classifier.

The authors Sethy A, Patra PK, Nayak S, Jena PM in 2018, used normalization and dilation to preprocess the NIT Rourkela dataset. From each of the 47 classes, only 150 specimens were taken. The discrete wavelet transform was used to extract features, which were then reduced using PCA. The feature set was provided to BPNN, and 94.8% accuracy was obtained. In 2019, the authors of implemented noise reduction, skew correction, and normalization towards preprocessing on the NIT Rourkela dataset. All 350 samples are taken from every 47 classes. The extraction of features was based on symmetric axis chords, mathematical features such as Euclidean distance and Hamilton distance, and the feature dimension was then reduced using PCA. These were applied to a Gaussian kernel with a radial basis function neural network and achieved a recognition rate of 98.8%.

Similarly various other techniques have been used for the same which might lengthen if mentioned.

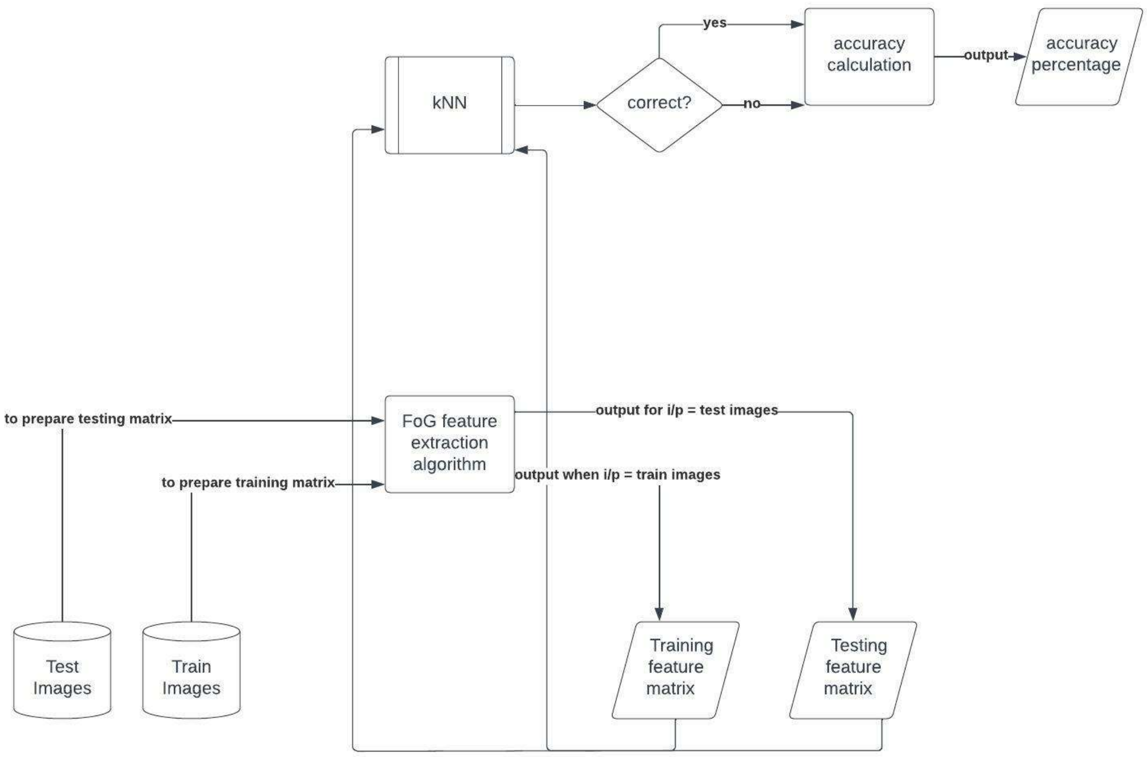
PROBLEM DEFINITION AND MOTIVATION

Offline OCR involves identification of letters from scanned textual images, and since no information is available about the movement of the marker, it happens to be more critical than online OCR.

Many of the Dravidian languages of India are not worked upon yet and hence no printed documents of old manuscripts are available on their languages. Many states of India also have some native languages on their course and hence it will be convenient if written documents can be made available online in printed form, hence, decrementing memory needed to store them and making them editable.

Many researchers have worked on the same and produced many algorithms to extract features from the handwritten characters to identify them using Machine Learning. Our goal was to test the efficiency of such algorithms and find their accuracy for future use.

DATA FLOW DIAGRAM


yes
accuracy calculation
output
accuracy percentage
kNN
correct?
no
FoG feature extraction algorithm
to prepare testing matrix
output for i/p = test images
to prepare training matrix
output when i/p = train images
Test Images
Train Images
Training feature matrix
Testing feature matrix

RESULTS

As the CTD algorithm for feature extraction was showing low accuracy on all of the modelling techniques such as LR, DT, kNN, GNB, RF, SVM, RNN and CNN, we decided to use only the FoG feature extraction algorithm and kNN to be our modelling technique, which is easy to implement and has a high accuracy score for classification. The dataset that we were provided was the IIITBOdiaV1 which consisted of 60 classes comprising of basic characters, and numbers and the accuracy on this dataset by the FoG features was 70.93%, and the claimed accuracy on the same using kNN modelling technique is 81.12%.

CONCLUSION

The Feature matrices extraction algorithm may produce high accuracy with the help of CNN classification algorithm, but yields less accuracy with primitive kNN. The number of columns (features) if increases, the prediction accuracy decreases and AMS feature matrix faces the worst consequence, although the

FoG matrix when input yields an accuracy of 70.93% on the dataset consisting of handwritten characters as well as numbers.

To yield high accuracy CNN must be used in case of AMS feature matrix or else the accuracy yields to be around 16% when the classification algorithm used in kNN on the dataset of 112 classes consisting of handwritten characters as well as numbers.

REFERENCE

- Research paper : Offline Odia handwritten character recognition with a focus on compound characters by Raghunath Dey, Rakesh Chandra Balabantaray and Sanghamitra Mohanty
- AnalyticsVidhya for machine learning
- Youtube for machine learning