Optical Character Recognition of Bengali Handwritten Words

MD Tanjim Mostafa, Rizvy Ahmed Kamal, Afreen Rahman Tithi, Dewan Golam Mortoza,
Ehsanur Rahman Rhythm,Md Sabbir Hossain, and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh

***Abstract*—Bengali Handwritten Word Recognition Using Optical Character Recognition is very important as the language is rich in terms of scripts and syntaxes. The effectiveness of deep learning methods for recognizing handwritten Bangla characters is investigated in the present paper. The creation and development of an OCR system for the Bangla language, including preprocessing techniques, feature extraction, and classification algorithms, is presented in this work. The suggested approaches are tested on a benchmark dataset to feed more accuracy, robustness, and computational efficiency.**

***Index Terms*—Bangla; Image Processing; Optical Character Recognition.**

I. Introduction

Optical character recognition (OCR) is the technique of automatically extracting text from images or scanned documents and converting it into machine-readable formats. OCR can be performed on digitally printed images or images with handwritten words. OCR systems are particularly interesting because they allow us to digitize traditional documents. Digital data has various advantages. For once, they are easy to manipulate which helps with the entry of new data and the modification of the existing document. Second, digital data help process the document, allowing us to perform tasks like search, context extraction, and translation of the document.

OCR systems can be used for automation purposes in libraries, post offices, educational institutes, etc. Because of its' usefulness, a lot of research has been done in several domains of OCR. Despite that, optimally and accurately recognizing Bangla handwritten characters still remain a daunting task because of how diverse and complicated the Bangla language is. Our paper explores different techniques of performing OCR on Bangla handwritten text to find out the optimal approach. [tentative]

II. Literature review

This section tries to show the relevant work basis on OCR for Bengali Handwritten word. There are many extraordinary work in OCR field have done by many researchers. Here are some related work mentioning.

Nadim Mahmud Dipu, Sifatul Alam Shohan proposed "Bangla OCR using Deep Learning based on Image classification Algorithm" suggested three different convolutional neural network based on the basis of image classification models which are trained and also examined on the BanglaLekha-Isolated dataset. The purpose is to recognise Bangla OCR from images. Among of three deep learning models VGG16 finds the highest rate of 98.93 percent. Moreover the model focuses in the construction of a state-of-the-art, which is acceptable in Bangla OCR system. By utilizing pre-trained deep learning models, it had implemented a transfer learning technique.

Farisa Bente Safir,Abu Quwsar Oni proposed another complete OCR that presents ent to ent OCR system that recognizes Bangla words from images. It can implement based on end to end architecture and it based on Bangla writting. The architecture based on four different pre-trained CNN architecture and uses two different bidirectional RNN. For improving the system a neural network is required rather than a pre-trained network.

THI TUYET HAI NGUYEN,ADAM JATOWT expressed an extensive survey to approaches post recorrecting printed OCRed texts which mainly English languages of Latin script. The author emphasis the importance of post OCR processing and analysis the the impact of OCR errors.

Using Convolutional Neural Network with data Augmentation for Bangla Hand-written Character recognition, Rumman Rashid Chowdhury, MD.Shahadat Hossain expresses the opportunity and methods for recognise the Bangla handwritten Character.CNN approaches is more convenient than machine learning process. The process finds the satisfying result for classifying to finf individual letters of the Bangla alphabet.

'Survey on OCR for Bangla and Devanagari scripts' the author proposed OCR in two major Indian scripts. It focuses on related to printed character and numerals,handwritten character/numerals. Here the author focuses the techniques which have been used in modern OCR systems and dataset used in OCR systems and for the better performance of OCR methods pre-processing methods being used.

III. Basics of Bangla Scripts

Bangla language consists of 50 letters in total. Among them, 11 are vowels and 39 are consonants. Two or more characters combine to form a new compound character called Juktoborno. Bangla characters are very diverse in shape which makes them difficult to deal with in OCR systems. There are 10 modifiers and 10 numeric characters as well. Bangla language has a

number of distinctive features, including the use of loops and hooks on some letters, also the use of diacritical marks above letters called 'matra' and diacritical marks below letters called 'hoshonto'. Some of these diacritical marks indicate distinct character pronunciation.

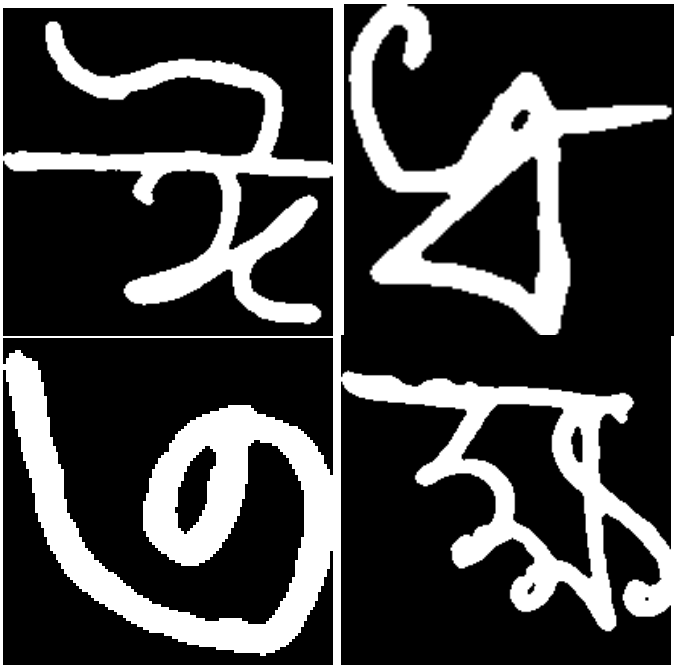

Fig. 1. Some Bangla characters

IV. Data Collection And Pre-processing

A. Data Collection

The dataset, BanglaLekha-Isolated was used in this experiment. It consists of handwriting samples of 50 Bangla basic characters, 10 Bangla numerals and 24 selected compound characters. There are 2000 handwriting samples for each of the 84 characters. The samples were collected from subjects of different age groups and gender.

V. Method

A. Understanding transformer based OCR

Generally, OCR systems consist of two main modules.

1) Text detection module
2) Text recognition module

Text detection module tries to detect the text blocks that exist in the source image either in word level or the text line level. This task can be thought of as an object detection problem, but instead of detecting objects in images, we are detecting text blocks in documents. The Text Recognition module is responsible for understanding the content of the detected text block and converting the visual signals into natural language tokens. The key difference between conventional Convolutional neural network(CNN) based OCR systems and transformer based OCR system is that the former models are built on top of existing CNN models for image understanding and recurrent neural network(RNN) for character level text generation whereas in the latter system a text line is passed to an image based encoder and then the output from the encoder is passed to a text based BERT-style decoder model.

1) Encoder: The source image is divided into small, equally-sized patches, treating them as individual elements similar to words in a sentence. Each patch is flattened into a one-dimensional vector and combined with positional embeddings. This embeddings are then passed through transformer encoder layers to process the image data. In OCR, the image is a series of localized text boxes. In order to maintain consistency across localized text boxes, the images or the text blocks on the images are reshaped to a uniform height and width (H x W). The image is then decomposed into patches of a specified size where the patch size is HW/(PxP). Here, P is the patch size. Each patch is then flattened and linearly projected to a D-Dimensional vector to create patch embeddings. These embeddings, together with two specialized tokens are then given learnable 1D position embeddings according to their absolute positions. Then the input sequence is passed through a series of encoder layers, each consisting of a multi-head self-attention module and a fully connected feed-forward network. After that, residual connections and layer normalization follows to create a more robust and stable representation of the input data.

References