A Review of Optical Character Recognition (OCR) Techniques on Bengali Scripts

Md Imdadul Haque Emon , Khondoker Nazia Iqbal , Md Humaion Kabir Mehedi ,
Mohammed Julfikar Ali Mahbub and Annajiat Alim Rasel

Department of Computer Science and Engineering
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
*{md.imdadul.haque.emon, khondoker.nazia.iqbal, humaion.kabir.mehedi,
mohammed.julfikar.ali.mahbub}@g.bracu.ac.bd*
annajiat@bracu.ac.bd

***Abstract*—The technique of transforming analogue documents into digital documents using document images is known as Optical Character Recognition (OCR). Since the mid-1980s, researchers have been studying Bangla character recognition [1]. The position of Bangla is seventh among the most popular languages [2]. OCR works for Bengali scripts have received a lot of attention but only few works have been done in this field because of the less availability of Bangla resources. Our paper provides an overview of the various OCR techniques for Bengali characters. We looked at a total of 13 papers and prepared a summary in this paper. It will help researchers to have a better understanding of the OCR techniques used in Bengali scripts. We hope that such a review will encourage researchers to continue working on and improving OCR for Bengali scripts.**

***Index Terms*—Optical Character Recognition (OCR), Bengali scripts, Digital documents**

I. Introduction

The process of transforming handwritten and scanned text documents into digital documents is known as OCR. In this technique, each character on a page is scanned individually. Optical Character Recognition started as a branch of pattern recognition, deep learning, and vision based research. An Austrian engineer developed the first OCR machine in the late 1920s [3] .

The most important technique for automatic text identification in today's world is optical character recognition (OCR). It has significantly enhanced the data entering procedure. We may make mistakes while copying data from an image to the computer. However, utilizing OCR technology, we can accurately transcribe picture documents into text documents. OCR also helps to save time and money by eliminating the need to manually enter data from scanned documents into a computer. In addition, we are unable to look for and change important information included in electronic images and pdf documents. OCR can help solve this problem by converting digital images and pdf files into modifiable and searchable digital text documents.

OCR technology has been used in several established languages throughout the world for many years. However, because Bangla is a low-resource language, OCR research in Bangla began in the late 1980s [1]. Bangla characters are more complex in form than other languages such as English, Arabic, French, etc. In addition, there are approximately 334 complex characters in the Bangla language. These characters are made up of many different of Bengali characters [4]. As a result, developing an OCR for Bangla is more complex than for other languages. For this reason, we have explored many techniques to implement OCR technology for the Bangla language in our study, which will be useful for future research in this field.

In Section II of this paper, we have summarized previously published work in this field. We gave an overview of the basic properties of Bengali scripts in Section III. Section IV contains information on the datasets utilized in various research work. The methodology and preprocessing stages then are outlined in Section V. Next, Section VI includes the result part. Finally, Section VI brings our paper to a conclusion.

II. Related Works

In this work, we analyzed around 13 research papers that investigated optical character recognition in the Bangla language. Various articles have used various models to develop OCR for the Bangla language. Hasnat et al.(2008) have presented an OCR system for detecting Bangla characters using the Hidden Markov Model in his research article [5]. The open source Tesseract engine was utilized for identifying Bangla scripts in another research [6]. Farisa et al. also constructed an OCR design employing different CNN architectures such as DenseNet, NasNet, and MobileNet in their research article [7]. To improve Bangla OCR output, some researchers used N-gram and Edit distance algorithms [8]. Neural networks are used as a better approach for detection of Bangla characters in a research paper [9]. Another research paper of Adnan et al., [10] used a Kohonen neural network to classify

Bangla characters in the OCR system . Back propagation neural network is used in the Shamim et al paper to recognize printed Bangla scripts [11]. Paul et al. [12] used a single hidden BLSTM framework for printed Bengali OCR systems.

The summary of our reviewed papers are listed in the below Table I

TABLE I
Related Work in OCR for Bangla Scripts

Paper	Technique	Dataset
Hasnat et al. [5]	HMM	Alphabets of Bangla Character set
Hasnat et al. [6]	Tesseract	All the characters of Bangla language.
Farisa et al. [7]	DenseNet, NasNet, MobileNet	Bengali Writing Dataset [13]
Sajib et al. [8]	N-gram and Edit Distance	83,570 words from different Online portals
Ahmed et al. [9]	Neural Network	Characters of Sutonny, Sulekha and Modhumati font
Adnan et al. [10]	Kohonen Network	Single character and single word image
Shamim et al [11]	Back Propagation Neural Network	Hundred thousand words from Bangla books, papers, and sixty thousand words from Bangla dictionary
Paul et al [12]	The BLSTM-CTC architecture	47,720 Text line images
Mahbub et al. [14]	Multilayer Feed Neural Network	Sutonny, Sulekha, Sunetra Fonts
Pal et al. [15]	A Feature based Tree classifier	Text documents from manually composed books containing 5000 characters in each

III. Properties of Bengali Scripts

Bangla language has the most complex-shaped characters than other languages. Bengali script is generally divided into two classes: vowels and consonants. There are 50 characters in all, including 11 vowels and 49 consonants [16]. There are also various modifiers and compound characters in Bengali scripts. Two or more consonants are combined to make compound characters. As the English characters, there are no upper and lower case letters for Bangla characters. But there is a new term in Bangla known as 'Matra' which is basically a horizontal line above the letters. Some characters are with 'Matra' and some characters have no 'Matra'. All these characteristics of Bangla alphabets are making very difficult to implement the OCR for Bangla. Some examples of Bangla alphabets are shown below in Table II .

TABLE II
Example of Bangla Characters

Vowels	অ, আ, ই, ঈ, উ, ঊ, ঋ, এ, ঐ, ও, ঔ
Consonants	ক, খ, গ, ঘ, ঙ, চ, ছ, জ, ঝ, ঞ, ট, ঠ, ড, ঢ, ণ, ত, থ, দ, ধ, ন, প, ফ, ব, ভ, ম, য, র, ল, হ, শ, ষ, স, য়, ড়, ঢ়, ৎ, ◌ং, ◌ঃ, ◌ঁ
Vowel Modifiers	◌া, ◌ি, ◌ী, ◌ু, ◌ূ, ◌ৃ, ◌ে, ◌ৈ, ◌ো, ◌ৌ
Compound Letters	ব্ব, জ্ঞ, ত্ত, ম্ন, ম্ল, গ্ধ, ক্ষ, ক্ল

IV. Dataset

Handwritten characters and machine printed images are two forms of OCR data for the Bangla language. Since Bangla is a low-resource language, the dataset is insufficient. Farisa et al. [7] in their paper used the Bangla Handwriting Dataset [13], which comprises 21,234 words gathered from the handwriting of 260 people. Some examples of the handwritten image data are shown in below Figure 1 and Figure 2. To train their model, some researchers utilized various Bangla font characters such as Sutonny, Sulekha, Modhumati, and Sunetra. Furthermore, Shamim et al. [11] have gathered hundred thousand words from Bangla publications and newspapers, as well as roughly sixty thousand words from the Bangla dictionary, for their research. The list of the datasets used in our reviewed papers are shown in the Table I.

V. Methodology

Different models were used in our reviewed publications to recognize Bangla characters in handwritten texts and scanned documents. CNN models (DenseNet121, NesNet, MobileNet), neural network models (Multilayer feed forward network, Back Propagation neural network, Kohonen Network), HMM, BLSTM, N-gram, and Edit Distance are some of the commonly used models in this domain.

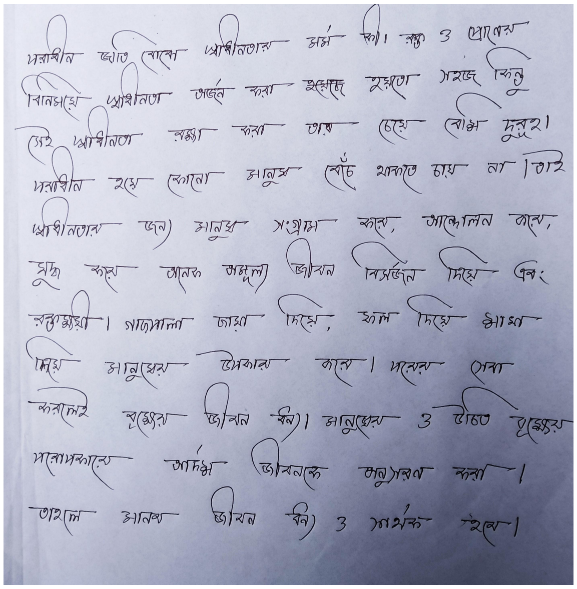
Although numerous approaches can be used to develop OCR technology, there are several basic processes that are followed in all 11 papers. The steps are as follows:

1) Preprocessing
2) Feature Extraction
3) Pattern Recognition and Classification
4) Post Processing

A. Preprocessing

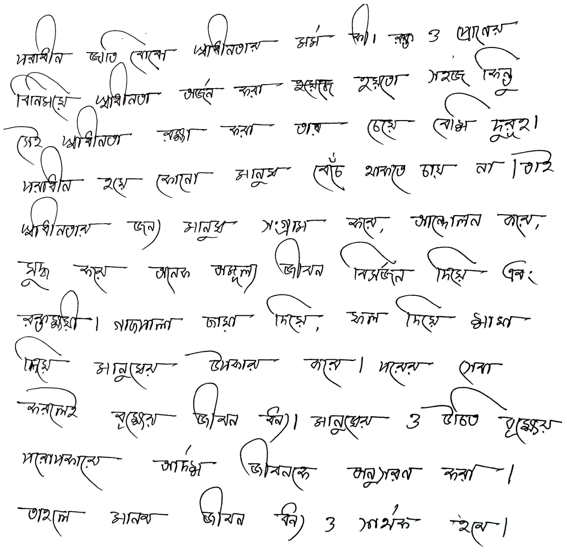
Some basic steps of preprocessing are described below:

- **Text Digitization and Binarization**:
  The popular techniques used in text digitization are Flat-bed scanner and hand-held scanner. Most of our reviewed papers, [5], [8], [14], [15] used Flat-bed scanner for digitization purpose. The technique of transforming gray scale photos to binary photos is known as Binarization. There are different Binarization techniques: Niblack's Algorithm, Otsu Global Algorithm, Global Fixed Threshold [17], etc. In some of our reviewed papers, Otsu Global algorithm is used and most of them used threshold approach for image Binarization.



পরাধীন জাতি বোঝে স্বাধীনতার মর্ম কী। রক্ত ও প্রাণের
বিনিময়ে স্বাধীনতা অর্জন করা হয়েছে হয়তো সহজ কিন্তু
সেই স্বাধীনতা রক্ষা করা তার চেয়ে বেশি দুরূহ।
পরাধীন হয়ে কোনো মানুষ বেঁচে থাকতে চায় না। তাই
স্বাধীনতার জন্য মানুষ সংগ্রাম করে, আন্দোলন করে,
যুদ্ধ করে আনন্দে অমূল্য জীবন বিসর্জন দিয়ে এবং
রক্তগঙ্গায়। গাছপালা ছায়া দিয়ে, ফল দিয়ে আশ্রয়
দিয়ে মানুষের উপকার করে। দশের সেবা
করলেই বৃক্ষের জীবন ধন্য। মানুষের ও উচিত বৃক্ষের
পরোপকারের আদর্শ জীবনকে অনুসরণ করা।
তাহলে মানব জীবন ধন্য ও সার্থক হবে।

Fig. 1. Example of Handwritten Raw Text Data



পরাধীন জাতি বোঝে স্বাধীনতার মর্ম কী। রক্ত ও প্রাণের
বিনিময়ে স্বাধীনতা অর্জন করা হয়েছে হয়তো সহজ কিন্তু
সেই স্বাধীনতা রক্ষা করা তার চেয়ে বেশি দুরূহ।
পরাধীন হয়ে কোনো মানুষ বেঁচে থাকতে চায় না। তাই
স্বাধীনতার জন্য মানুষ সংগ্রাম করে, আন্দোলন করে,
যুদ্ধ করে আনন্দে অমূল্য জীবন বিসর্জন দিয়ে এবং
রক্তগঙ্গায়। গাছপালা ছায়া দিয়ে, ফল দিয়ে আশ্রয়
দিয়ে মানুষের উপকার করে। দশের সেবা
করলেই বৃক্ষের জীবন ধন্য। মানুষের ও উচিত বৃক্ষের
পরোপকারের আদর্শ জীবনকে অনুসরণ করা।
তাহলে মানব জীবন ধন্য ও সার্থক হবে।

Fig. 2. Example of Handwritten Binary Data

- **Noise Removal**:
  Noise removal is an essential step of data precoressing. While scanning documents or images, it can be corrupted by noise. The noises should be removed before data processing to recognize Bangla characters accurately. The most common noises for the printed and scanned documents are the background noises and the salt and pepper noises. According to our analysed papers, Low-pass filters, wiener filters and median filters are the most commonly used techniques for noise removal.
- **Skew Detection and Correction**:
  Papers can be skewed during scanning with different scanners, such as a hand-held scanner or a flat-bed scanner, therefore it is required to align the documents accurately before sending them for training. The skewness determination technique for Bangla is based on the existence of 'Matra'. The following are the basic two phases in removing skewness from a dataset:
  1) Calculate the value of skew angle Θt
  2) Image rotation in the opposite direction of the skew angle Θt
- **Segmentation**:
  Various forms of segmentation are carried out during preprocessing. Lines, words, and characters can all be segmented. Greedy search technique and Contour tracing techniques are used in the character segmentation process.

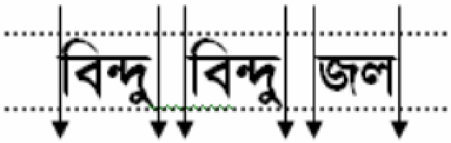


Fig. 3. Word Segmentation from Printed Documents [18]

B. Feature Extraction

We need to extract the features after the preprocessing. Each character must be represented as a vector in order to extract the features. To extract features, Mahbub et al.[14] applied scaled character and the Freeman chain code in their research. Hasnat et al. [5], on the other hand, separated the preprocessed images into distinct frames of a predetermined length. Then they calculated Discrete Cosine Transform for each pixel in the frames. To extract features from images, some publications applied the DFS approach, while others used the RNN model. In their study, Shamim et al. [11] focused on distinguishing connected components in a character in order to extract properties from the input images. They extracted connected components from each character using the DFS method. In their research, Farisa et al. [7] introduced two processes for extracting features from an input image. Initially, a baseline model is utilized, which accepts images as input and produces high-dimensional characteristics as output. CNN based models: DenseNet, NasNet, MobileNet are used to build their baseline model. The output of the baseline model is then supplied as input to a bidirectional RNN model in the second stage. To tackle the gradient problem of the RNN model, they applied LSTM and Gated Recurrent Unit [19]. Finally, the output of the RNN model is used by additional models for pattern recognition.

C. Pattern Recognition and Classification

The key aspect of OCR for Bangla is pattern recognition and categorization. All extracted characteristics are compared to the training dataset in this stage to identify and categorize Bangla characters. In a paper [14], a multi-layer feed forward is utilized to detect and categorize Bangla letter patterns. A temporary model and a Hidden Markov Model Toolkit (HTK) recognizer are used to discover patterns in another work [5]. Back propagation recognizer, a common multi-layer network learning approach, has been applied in the research work of [11] to solve pattern categorization problems. Tesseract was utilized in Hasnat et al. [6] paper to identify patterns and produce text data from OCR. To categorize Bangla characters, a paper [12] for the printed Bengali OCR system, used 166 class labels. Adnan et al. [7] utilized the Kohonen network to identify patterns. In a study [15], a template machine approach is developed for categorization problems. Furthermore, To identify and recognize Bangla characters, Ahmed et al. [9] utilized a Neural network.

D. Post Processing

The OCR system's final phase is post-processing. This stage is carried out once the texts have been identified and categorised. The recognized texts don't always match the source texts. Various post-processing stages, including as error checking, spell checking, and text editing, are used to remedy these mistakes. In their study, Hasnat et al. [5] developed a suggestion-based spell checker to fix textual mistakes. Rather than replacing words with inaccuracies, they use a different technique to suggest other ways to correct those words. In the post processing step, edit distance technique along with N-gram algorithms are used by the authors of a work on improving Bangla OCR output [8].

VI. Result Analysis

The articles we looked at used a variety of techniques to implement a Bengali OCR system. The character level accuracy rate and error rate are displayed in Table III and Table IV.

TABLE III
Accuracy Score

Model	**Accuracy**
Multilayer Feed Forward Network	97%
HMM	98%
Back Propagation	97.5%
Tesseract	93%
Neural Network	96.33%
Kohonen Network	99%
Tree Classifier	96.55%
BLSTM-CTC	**99.32%**

TABLE IV
Error Rate

Model	**Error Rate**
DenseNet121+GRU	0.091
Edit Distance	0.1685
N-gram approach	0.1716

We can conclude from the accuracy level that BLSTM-CTC architecture obtained the highest accuracy when it comes to develop OCR for the Bangla language.

VII. Conclusion

In this paper we reported various works on OCR for Bengali scripts. We looked at a total of 13 papers, 10 of which applied different approaches to develop an OCR and 3 of them provided an overview of Bangla OCR. We discovered through our research that several types of Neural Networks are commonly utilized in Bengali character identification and categorization. Furthermore, the BLSTM network achieved the highest accuracy in Bengali character recognition, with a 99.32% accuracy, according to our research. Our survey will encourage the researchers to work on Bangla OCR . It will help the researchers to build an OCR for Bangla, which will be able to detect and classify Bengali characters from both handwritten and scanned documents.

References

[1] T. Ahmed, M. N. Raihan, R. Kushol, and M. S. Salekin, "A complete bangla optical character recognition system: An effective approach," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 2019, pp. 1–7. doi: 10.1109/ICCIT48885.2019.9038551.

[2] M. Hasan, *Bangla ranked at 7th among 100 most spoken languages worldwide*, 2020. [Online]. Available: https://archive.dhakatribune.com/world/2020/02/17/bengali-ranked-at-7th-among-100-most-spoken-languages-worldwide (visited on 12/10/2021).

[3] *Optical character recognition – history of optical character recognition (ocr)*, 2021. [Online]. Available: https://history-computer.com/optical-character-recognition-history-of-optical-character-recognition-ocr/ (visited on 12/12/2021).

[4] N. Das, K. Acharya, R. Sarkar, S. Basu, M. Kundu, and M. Nasipuri, “A benchmark image database of isolated bangla handwritten compound characters,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, pp. 413–431, Dec. 2014. DOI: 10.1007/s10032-014-0222-y.

[5] M. A. Hasnat, S. M. Habib, and M. Khan, “A high performance domain specific ocr for bangla script,” in *Novel algorithms and techniques in telecommunications, automation and industrial electronics*, Springer, 2008, pp. 174–178.

[6] M. A. Hasnat, M. R. Chowdhury, and M. Khan, “An open source tesseract based optical character recognizer for bangla script,” in *2009 10th International Conference on Document Analysis and Recognition*, IEEE, 2009, pp. 671–675.

[7] F. B. Safir, A. Q. Ohi, M. F. Mridha, M. M. Monowar, and M. A. Hamid, “End-to-end optical character recognition for bengali handwritten words,” in *2021 National Computing Colleges Conference (NCCC)*, IEEE, 2021, pp. 1–7.

[8] M. S. Ahmed, T. Gonçalves, and H. Sarwar, “Improving bangla ocr output through correction algorithms,” in *2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, IEEE, 2016, pp. 338–343.

[9] A. A. Chowdhury, E. Ahmed, S. Ahmed, S. Hossain, and C. M. Rahman, “Optical character recognition of bangla characters using neural network: A better approach,” in *2nd ICEE*, Citeseer, 2002.

[10] A. M. Shatil *et al.*, “Research report on bangla optical character recognition using kohonen network,” BRAC University, Tech. Rep., 2007.

[11] S. Ahmed, A. N. Sakib, M. Ishtiaque Mahmud, H. Belali, and S. Rahman, “The anatomy of bangla ocr system for printed texts using back propagation neural network,” *Global Journal of Computer Science and Technology*, 2012.

[12] D. Paul and B. B. Chaudhuri, “A blstm network for printed bengali ocr system with high accuracy,” *arXiv preprint arXiv:1908.08674*, 2019.

[13] M. Biswas, R. Islam, G. K. Shom, *et al.*, “Banglalekha-isolated: A multi-purpose comprehensive dataset of handwritten bangla isolated characters,” *Data in brief*, vol. 12, pp. 103–107, 2017.

[14] M. M. Alam and M. A. Kashem, “A complete bangla ocr system for printed characters,” *JCIT*, vol. 1, no. 01, pp. 30–35, 2010.

[15] U. Pal and B. Chaudhuri, “Ocr in bangla: An indo-bangladeshi language,” in *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, IEEE, vol. 2, 1994, pp. 269–273.

[16] *Bengali alphabet - wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Bengali_alphabet.

[17] M. G. Kibria and Al-Imtiaz, “Bengali optical character recognition using self organizing map,” in *2012 International Conference on Informatics, Electronics Vision (ICIEV)*, 2012, pp. 764–769. DOI: 10.1109/ICIEV.2012.6317479.

[18] F. Y. Omee, S. S. Himel, M. Bikas, and A. Naser, “A complete workflow for development of bangla ocr,” *arXiv preprint arXiv:1204.1198*, 2012.

[19] *Gated recurrent unit - wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Gated_recurrent_unit.