Semiconductor Optoelectronics
ISSN: 1001-5868

ESTIMATING OUTCOMES OF TEXT LINE SEGMENTATION AND SKEW ESTIMATION IN HANDWRITTEN DOCUMENTS OF BANGLA LANGUAGE

Prosenjit Mukherjee[1], Dr. Akash Saxena[2]
[1]*Research Scholar, Department of Computer Science & Engineering, Mansarovar Global University, Sehore, M.P., India.*
[2] *Research Guide, Department of Computer Science & Engineering, Mansarovar Global University, Sehore, M.P., India.*

ABSTRACT
Because of its complex structure, marked by ligatures, conjuncts, and diacritic markings, the Bangla script is notoriously difficult to distinguish. Also, the operation is complicated by variations in writing styles, stroke thickness, and image quality. The proposed approach uses the handwritten Bangla paper as the dataset for improvised text-line segmentation and skew estimation. Filtering, grayscale conversion, and binarization are the three preprocessing techniques that can be used. Distances between text lines are being evaluated using the ESLD (Enhanced Supervised Learning Distance) method, while words or Connected Components are being clustered using G-Clustering. Skew estimate can also be done by determining the skew angle in relation to the void.

Keywords: Bangla, Image, Words Segmentation, Skew, Algorithm

I. INTRODUCTION

The study of developing methods and systems for automatically detecting and understanding handwritten letters is known as Handwritten Character Recognition (HCR) or Optical Character Recognition (OCR). In recent years, this technology has received a lot of interest because of all the many fields it may be used in, such as digitizing historical documents, enhancing data entry processes, enabling automated form processing, and improving human-computer interaction.

Handwritten character recognition is a complex task that involves several stages of processing, including image acquisition, preprocessing, feature extraction, and classification. The ultimate goal is to convert handwritten text into machine-readable format, allowing computers to understand and interpret the content accurately.

The process begins with image acquisition, where the handwritten document is captured using a scanner or a digital camera. This step is critical, as the quality of the acquired image significantly affects the accuracy of the subsequent recognition process. Factors such as resolution, lighting conditions, and document alignment play a crucial role in ensuring optimal image quality.

Once the image is obtained, the next step is preprocessing. Preprocessing techniques are employed to enhance the quality of the image and improve the recognition performance. Common preprocessing steps include noise removal, normalization, binarization, and

segmentation. Noise removal techniques eliminate unwanted artifacts and smoothen the image, while normalization aims to standardize the size, orientation, and slant of the characters. Binarization converts the grayscale image into a binary image, where pixels are classified as either foreground or background, simplifying subsequent processing steps. Segmentation involves separating individual characters from the text, which can be challenging due to variations in writing styles, overlapping characters, or connected strokes.

Handwritten character recognition has numerous practical applications. In the field of document digitization, HCR enables the conversion of handwritten text into editable and searchable electronic formats, facilitating information retrieval and document analysis. HCR also plays a crucial role in automated form processing, where handwritten forms can be processed rapidly, eliminating the need for manual data entry. In addition, HCR has applications in personal devices, such as tablets and smartphones, allowing users to input text through handwriting recognition rather than traditional keyboards.

II. REVIEW OF LITERATURE

Safir, Farisa et al., (2021) The process of digitizing paper documents is known as optical character recognition (OCR). Multiple commercial and non-commercial OCR systems for printed and handwritten copies of a variety of languages are now available. Despite this, resources for recognizing Bengali words are quite limited. The majority of these studies included optical character recognition (OCR) of printed Bengali characters. Complete OCR for the Bengali language is presented in this publication. The suggested architecture is a comprehensive solution for recognizing photographs of handwritten Bengali text. We construct the OCR design by experimenting with well-known convolutional neural network (CNN) architectures as DenseNet, Xception, NASNet, and MobileNet. In addition, we test out two distinct RNN approaches, namely LSTM and GRU. We put the suggested architecture through its paces on the Bengali handwritten image dataset BanglaWritting, which has undergone rigorous scholarly assessment. Using a DenseNet121 model with a GRU recurrent layer, the suggested technique obtains an error rate of 0.091 per character and 0.273 per word.

Isthiaq, Asif & Saif, Najoa (2020) In the computing industry, the term "optical character recognition" has become trendy. For quite some time now, people have been using artificial neural networks for character recognition. Due to the non-linear and complicated nature of many real-world interactions between inputs and outputs, ANN's capacity to learn and simulate such relationships is crucial. Compared to the English language, research on OCR with the Bangla language is limited. Therefore, there is room for investigation here. In a matter of seconds, you may search and scan through hundreds of Bangla documents and quickly edit the data. For those with visual impairments, for example, OCR software may assist convert written materials like books, periodicals, and newspapers into audio files that can be played on a computer or portable audio player. Traditional OCR suffers from not having access to a large enough dataset, not having access to every possible font of characters, and having trouble accurately recognizing a large number of complicated and similarly shaped letters. Since neural networks require a lot of data to train, we began our investigation by trying to collect enough information to create a sufficiently sized dataset. We collected our own data consisting of 2,97,898 photos of individual Bangla characters from a variety of typefaces. Then, we utilized the multi-layer perceptron classifier from Scikit-learn to create a neural network, and we built

our own multi-layer feed forward back propagation neural network from scratch in the Tensorflow machine learning framework. We've also developed a graphical user interface to show how well our system can recognize photos of single Bangla characters.

Ghosh, Tapotosh et al., (2020) The Bangla script is a complicated alphabet, making handwritten character identification a challenging process. OCR models should be mobile-friendly because of the widespread use of OCR on smartphones and tablets. There have been a lot of studies done on this topic, but none of them have been able to identify more than 200 characters with any kind of reliability. MobileNet is a state-of-the-art CNN architecture optimized for small devices like smartphones and tablets. In this research, MobileNet was utilized for OCR (optical character recognition). It has a 96.17% success rate in identifying 171 compound character classes, a 98.37% success rate in identifying 50 basic character classes, and a 99.56% success rate in identifying 10 numerical character classes.

Rizvi, Md. Atiqul Islam et al., (2019) Recent research on recognizing handwritten Bangla characters has received a lot of interest. Numerous different types of characters, including numbers, basic characters, compound characters, and modifier characters, are used in the Bangla language. Recognizing handwritten text is difficult because of the complexity of cursive writing and the inherent variance across writers. This research takes a look at two distinct methods for recognizing handwritten Bangla characters and compares their respective results. In the first, a classifier-based approach, features are extracted using a hybrid model of the feature extraction technique, and recognition is carried out by a multiclass support vector machine (SVM). The second one utilizes a CNN for its computations. Ten Bangla numbers, fifty basic characters, and a subset of complex characters often used in Bangla were taken into account for recognition. The experimental findings show that the CNN model is superior to the conventional classifier-based method, with recognition rates of 98.04 percent for basic Bangla letters, 99.68 percent for numbers, and 98.18 percent for the subset of compound characters.

Omee, Farjana et al., (2011) A large number of algorithms and techniques are needed to create a Bangla OCR. A great amount of work was put into creating a Bangla OCR. None of them, however, managed to produce a flawless Bangla OCR. There are defects in every one of them. We spoke about the range of issues with present Bangla OCR software. In this paper, we detail the foundational requirements for creating an OCR for the Bangla language and provide a comprehensive methodology for doing so, naming all the viable algorithms along the way.

III. RESEARCH METHODOLOGY

The study introduces a novel technique for estimating the skew of Handwritten Bangla documents and segmenting their text lines. In order to calculate the skew for the crooked text lines, the proposed technique uses the ESLD (Enhanced Supervised Learning Distance) and R-Clustering algorithms to detect and classify handwritten text lines. Gap estimation, or the process of determining how much space there is between any given pair of lines of text or words, is carried out via the ESLD (Enhanced Supervised Learning Distance) method. Next, the R-Clustering method is used to classify the interdependent parts into clusters. Each method entails four steps: (i) document picture input, (ii) preprocessing, (iii) text line segmentation, and (iv) skew estimate.

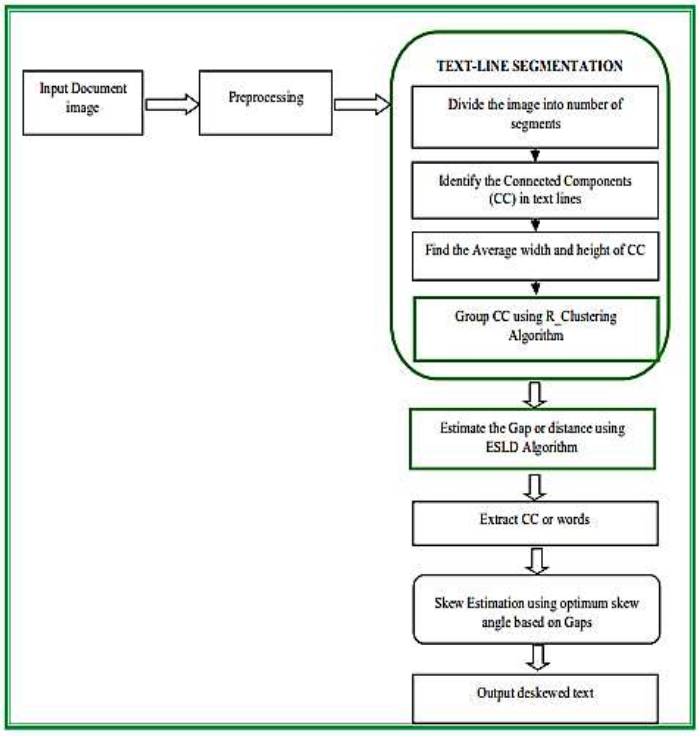

Figure 1: Overall Proposed Segmentation Architecture

Preprocessing

As an input, the picture of the handwritten document is cleaned up. Filtering, grayscale conversion, and binarization are the three methods used in preprocessing. This procedure guarantees the best possible segmentation precision. The rows and columns of the picture received represent the document's layout.

Text line Segmentation

The handwritten texts are first separated into lines, and then the words inside those lines are segmented by identifying the CCs (Connected Components). After figuring out the typical size of all the parts, we get Connected Components.We utilized these algorithms:

- Find the Average width and height of CC
- Group CC using R-Clustering Algorithm
- Connected Components (CC) using R-Clustering Algorithm
- Estimate the Gap or distance using SLD Algorithm
- Supervised Learning Distance (SLD) Algorithm
- Extraction of Connected components (CC)

Skew Estimation using Optimum Skew Angle

Using this procedure, the connected components (CCs) are isolated and their optimal skew angles are calculated. The document's picture is initially sent through an Edge Detector for preprocessing and edge extraction. Then, to find the interconnected parts, we dilate the extracted edges using a circular structuring element.

IV. RESULTS AND DISCUSSION

Images of handwritten Bangla documents with a total of 262 lines and 820 words are used to test the suggested deskewing method. Statistically, we may expect a success rate of 98.5% for line segmentation and 93.4% for word segmentation. For segmenting crooked lines and words, it is made clear that the suggested technique achieves competitive performance. The SLD algorithm is used to successfully accomplish line segmentation in this approach. Degradation in word segmentation performance is shown as a result of irregular spacing between words and broken letters. Word and line segmentation results are summarized in Table 1.

Table 1: Performance of the Proposed Segmentation

Approaches	Text lines	Words	Accuracy
Input	262	820	98.5%
Proposed Segmentation	255	765	94.25%

Figure 2 shows the outcome of segmenting the text lines.

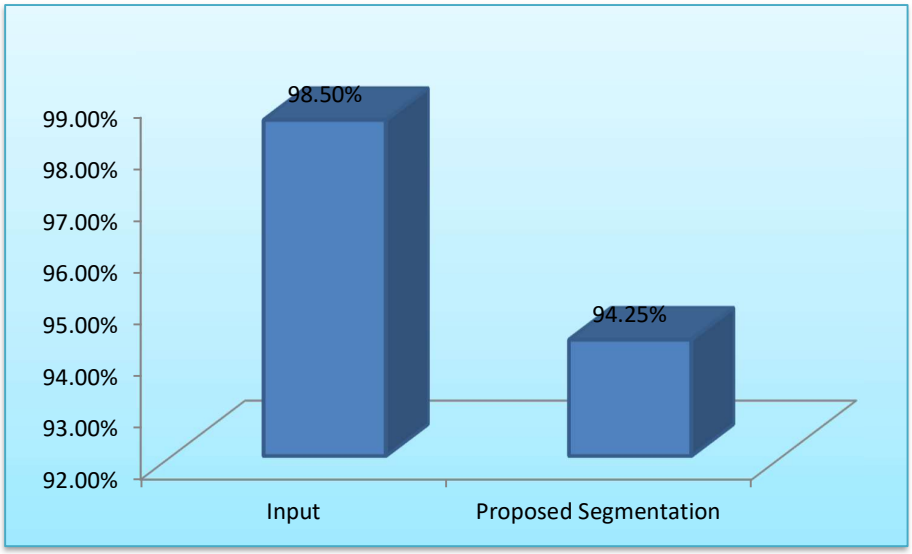

Figure 2: Analysis of Line and Word Segmentation Using a proposed Approach

The suggested algorithm has been tested on handwritten text that has been collected from people of varying educational and socioeconomic backgrounds. About 4800 handwritten Bangla words are used in the experiment, with 40% used to train the classifier and the remaining 20% used for testing. About 2,500 Bangla words were examined, with a success rate of 97.25 percent. In Table 2, we can see the classifier's skew correction accuracy.

Table 2: Skew Correction Algorithm Outcomes

Total Words	Skewed words	Corrected	Incorrected	Skew Correction Accuracy
4800	2500	3150	92	97.25%

Distance metrics are learned and tested on a subset of 100 documents (5 total), containing a total of 1106 text lines. Table 3 shows the true detection rates of text lines using the SLD method with and without the metric learning.

Table 3: SLD-corrected line detection rates in texts

Different Metric	Detected Text Lines
With learned metric	96.00%
With metric by hand	88.00%

Using distance metric learning has been shown to significantly improve text line segmentation performance. Despite the fact that the SLD algorithm and metric learning function well, there are still some text line identification failures. There are two possible types of failures here: When an actual text line is split into two or more lines (representing several clusters), this is called error line splitting; when two or more real text lines are merged into one cluster, this is called error line merging.

V. CONCLUSION

The suggested approach can accurately ascertain the optimal angle, deskew the word, and store the words on the line in the most efficient manner possible. Line and word segmentation techniques show comparable effectiveness despite differences in character size and the presence of consonant and vowel modifiers. Promising results have been shown using the suggested methodologies to reliably segment text lines and estimate skew, opening the door to higher recognition accuracy and more uses in the field of Bangla handwriting recognition. To further improve the performance of recognition algorithms, it would be helpful to have access to large-scale annotated datasets that are tailored to Bangla handwriting during training and assessment.

REFERENCES: -

1. Safir, Farisa & Ohi, Abu & Ph. D., M. & Monowar, Muhammad Mostafa & Hamid, Md. Abdul. (2021). End-to-End Optical Character Recognition for Bengali Handwritten Words.

2. Shakunthala b s, & dr. C s pillai. (2021). Enhanced text line segmentation and skew estimation for handwritten kannada document. Journal of Theoretical and Applied Information Technology. 15th January 2021. Vol.99. No 1. ISSN: 1992-8645.

3. Isthiaq, Asif & Saif, Najoa. (2020). OCR for Printed Bangla Characters Using Neural Network. International Journal of Modern Education and Computer Science. 12. 19-29. 10.5815/ijmecs.2020.02.03.

4. Ghosh, Tapotosh & Abedin, Md & Chowdhury, Shayer & Tasnim, Zarin & Karim, Tajbia & Reza, S M Salim & Saika, Sabrina & Yousuf, Mohammad. (2020). Bangla handwritten character recognition using MobileNet V1 architecture. Bulletin of Electrical Engineering and Informatics. 9. 2547-2554. 10.11591/eei.v9i6.2234.

5. Rizvi, Md. Atiqul Islam & Kaushik, Deb & Khan, Mohammad & Kowsar, Mir & Khanam, Tahmina. (2019). A comparative study on handwritten Bangla character recognition. Turkish Journal of Electrical Engineering and Computer Sciences. 27. 3195-3207. 10.3906/elk-1901-48.

6. Rakshit, Payel & Halder, Chayan & Roy, Kaushik. (2019). An Approach toward Character Recognition of Bangla Handwritten Isolated Characters. 10.1201/9780429277573-2.

7. Omee, Farjana & Himel, Shiam & Bikas, Md. Abu Naser. (2011). A Complete Workflow for Development of Bangla OCR. International Journal of Computer Applications. 21. 1-6. 10.5120/2543-3483.

8. Bag, Soumen & Bhowmick, Partha & Harit, Gaurav & Biswas, Arindam. (2011). Character Segmentation of Handwritten Bangla Text by Vertex Characterization of Isothetic Covers. 247-250. 10.1109/NCVPRIPG.2011.12.

9. Basu, Subhadip & Das, Nibaran & Sarkar, Ram & Kundu, Mahantapas & Nasipuri, Mita & Basu, Dipak. (2009). A hierarchical approach to recognition of handwritten Bangla characters. Pattern Recognition. 42. 1467-1484. 10.1016/j.patcog.2009.01.008.