Challenges And Solutions In Ocr Of Handwritten Bangla Documents

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

Optical character recognition (OCR) is a process of converting analogue documents into digital using document images. Currently, many commercial and non-commercial OCR systems exist for both handwritten and printed copies for different languages. Despite this, very few works are available in case of recognizing Bengali words. Among them, most of the works focused on OCR of printed Bengali characters. While OCR for printed texts has made significant progress, the OCR of handwritten documents poses unique challenges, particularly for scripts with complex character shapes and variations, such as Bangla. This research paper explores the issues related to OCR of handwritten Bangla documents and proposes potential solutions to improve the accuracy and efficiency of the recognition process.

Keywords: Optical Character Recognition, Script, Bengali, Ligatures, Writing.

I. INTRODUCTION

Optical Character Recognition (OCR) is a technology that plays a crucial role in converting scanned or photographed documents into editable and searchable formats. It has greatly facilitated document digitization, information retrieval, and data analysis across various domains. While OCR systems have achieved significant advancements in recognizing printed texts, the accurate recognition of handwritten documents remains a challenging task. This challenge becomes particularly pronounced when dealing with scripts that have complex character shapes and variations, such as Bangla.

Bangla, also known as Bengali, is the official language of Bangladesh and the second most widely spoken language in India. It has a rich and diverse script with a distinct set of characters, ligatures, and stroke patterns. The handwritten form of Bangla script adds an additional layer of complexity, making it difficult for conventional OCR systems to accurately interpret and transcribe the text. Therefore, developing effective OCR solutions for handwritten Bangla documents is of paramount importance to unlock their digital potential, preserve cultural heritage, and facilitate language processing applications.

Handwritten Bangla script exhibits several distinctive characteristics that pose challenges for OCR systems. Firstly, the shape and form of characters can vary significantly depending on the writer's style, resulting in a wide range of character variations. Secondly, Bangla script incorporates ligatures, which are combinations of two or more characters that form a single unit. These ligatures introduce additional complexities in the recognition process, as they require accurate segmentation and interpretation. Furthermore, cursive writing styles in Bangla can lead to overlapping strokes, making it challenging to separate individual characters. The presence of noise, smudges, and irregularities in handwritten texts further exacerbates the difficulty of accurate recognition. Finally, the limited availability of annotated training datasets specifically for handwritten Bangla poses a significant challenge for training robust and accurate OCR models.

Although various OCR systems exist for recognizing printed Bangla texts, their performance on handwritten documents is far from satisfactory. Many existing OCR techniques are designed primarily for Latin-based scripts, and their adaptability to non-Latin scripts like Bangla is limited. These systems often struggle to accurately segment characters, handle ligatures, and deal with the inherent complexities of handwritten Bangla script. Therefore, there is a need for specialized OCR solutions tailored to the unique characteristics of handwritten Bangla script.

Preprocessing techniques, such as noise reduction and enhancement, can help improve the quality of handwritten documents and enhance the effectiveness of subsequent recognition processes. Character segmentation algorithms specifically designed for Bangla script can aid in accurately separating individual characters and ligatures. Additionally, feature extraction methods can capture the distinctive characteristics of handwritten Bangla script, facilitating accurate recognition. Machine learning and deep learning approaches can be employed to train models that can effectively learn and recognize handwritten Bangla script.

II.OCR TOOLS AND TECHNOLOGIES

There are several OCR tools and technologies available that can be utilized for various OCR applications, including the recognition of handwritten Bangla documents. Here is an overview of some widely used OCR tools and technologies:

Tesseract OCR

Tesseract is one of the most popular open-source OCR engines developed by Google. It supports a wide range of languages, including Bangla, and provides high accuracy in recognizing printed text. Tesseract can be customized and trained for specific OCR tasks, making it a versatile tool.

ABBYY FineReader

ABBYY FineReader is commercial OCR software that offers advanced OCR capabilities for both printed and handwritten texts. It provides excellent accuracy and supports multiple languages, including Bangla. ABBYY FineReader offers features such as layout analysis, text extraction, and document conversion.

Adobe Acrobat Pro

Adobe Acrobat Pro is a comprehensive software suite that includes OCR functionality. It allows users to perform OCR on scanned documents and images, converting them into searchable and editable formats. Adobe Acrobat Pro supports multiple languages, making it suitable for recognizing Bangla text.

Microsoft Azure OCR

Microsoft Azure OCR is a cloud-based OCR service provided by Microsoft Azure Cognitive Services. It offers powerful OCR capabilities for printed and handwritten text recognition. It supports multiple languages, including Bangla, and provides APIs for easy integration into applications.

Amazon Textract

Amazon Textract is a cloud-based OCR service provided by Amazon Web Services (AWS). It is designed to extract text and data from scanned documents, forms, and images. Amazon Textract utilizes machine learning algorithms to achieve accurate OCR results and supports multiple languages, including Bangla.

Open OCR

OpenOCR is an open-source OCR framework that supports various OCR engines, including Tesseract. It provides a flexible and customizable platform for OCR development and integration into different applications. OpenOCR can be extended and adapted for specific OCR requirements.

Google Cloud Vision OCR

Google Cloud Vision OCR is a cloud-based OCR service offered by Google Cloud. It provides accurate OCR capabilities for both printed and handwritten text recognition. Google Cloud Vision OCR supports multiple languages, including Bangla, and offers APIs for easy integration.

Pytesseract

Pytesseract is a Python wrapper for Tesseract OCR engine. It allows developers to utilize the OCR functionality of Tesseract within Python applications. Pytesseract provides a simple and convenient interface for integrating OCR capabilities into Python projects.

These OCR tools and technologies offer a range of features and functionalities for recognizing and extracting text from various types of documents, including handwritten Bangla documents. They provide options for customization, language support, and integration into different platforms, enabling developers and users to implement OCR solutions according to their specific requirements.

III.CHALLENGES IN OCR OF HANDWRITTEN BANGLA DOCUMENTS

OCR of handwritten Bangla documents poses several challenges due to the complex nature of the script and variations in writing styles. These challenges affect the accuracy and reliability of recognition algorithms. The key challenges in OCR of handwritten Bangla documents are as follows:

Character Segmentation

Accurate segmentation of individual characters is a fundamental step in OCR. However, handwritten Bangla script often features connected and overlapping strokes, making it difficult to separate characters. Ligatures, which are combinations of characters, further complicate the segmentation process. OCR systems need to develop robust algorithms that can effectively identify and segment individual characters and ligatures in handwritten Bangla script.

Variability in Character Shapes

Handwritten Bangla characters exhibit significant variability in their shapes and forms. Different writers may have their distinct style, resulting in variations in character proportions, stroke placement, and connecting strokes. OCR systems must be able to handle these shape variations and account for the wide range of possible representations for each character.

Ligature Recognition

Bangla script extensively uses ligatures, where two or more characters are combined to form a single unit. Ligatures pose a challenge in OCR as they require accurate identification, segmentation, and recognition of constituent characters. Handling the different ligature variations and determining the appropriate combination of characters within a ligature is crucial for accurate recognition of handwritten Bangla script.

Cursive Writing Styles

Handwritten Bangla script often involves cursive writing, where characters are connected with fluid strokes. The cursive nature of the script leads to overlapping strokes, making it challenging to distinguish individual characters. OCR systems must employ sophisticated algorithms to separate and recognize characters correctly, considering the cursive writing style and the continuity of strokes.

Noise and Irregularities

Handwritten documents are susceptible to noise, smudges, and irregularities caused by imperfect writing surfaces, ink spread, or unintended marks. These artifacts can distort the appearance of characters, making it difficult for OCR systems to accurately interpret the text. Robust preprocessing techniques are necessary to reduce noise, enhance image quality, and improve the reliability of character recognition.

Limited Annotated Training Data

OCR systems heavily rely on large amounts of accurately annotated training data for effective training and recognition. However, for handwritten Bangla script, the availability of such annotated datasets is limited compared to printed Bangla. The scarcity of training data hinders the development of robust and accurate OCR models. Addressing this challenge requires efforts to create comprehensive and diverse annotated datasets specifically for handwritten Bangla script.

Overcoming these challenges requires the development of specialized algorithms and techniques that can handle the unique characteristics of handwritten Bangla script. Advanced segmentation algorithms, shape normalization techniques, ligature recognition models, and robust preprocessing methods can contribute to improving the accuracy and reliability of OCR systems for handwritten Bangla documents. Additionally, the availability of larger annotated datasets and collaborations with language experts and writers can help address the challenges and enhance the performance of OCR for handwritten Bangla script.

IV.SOLUTIONS AND METHODOLOGIES

To address the challenges in OCR of handwritten Bangla documents, various solutions and methodologies can be employed. The proposed approaches aim to improve accuracy, enhance segmentation, handle ligatures, and mitigate the impact of noise and irregularities. The following are some key solutions and methodologies:

Preprocessing Techniques

- Noise Reduction: Employ image processing techniques such as Gaussian filtering, median filtering, or adaptive filtering to reduce noise and smudges in handwritten Bangla documents.

- Image Enhancement: Apply contrast adjustment, histogram equalization, or adaptive enhancement methods to improve the legibility of characters and enhance the overall quality of the handwritten document.

Character Segmentation

- Stroke Analysis: Utilize stroke analysis algorithms to identify and separate individual strokes in handwritten Bangla characters.

- Curve Fitting: Apply curve fitting techniques, such as spline interpolation, to approximate and segment characters based on stroke paths.

- Connected Component Analysis: Utilize connected component analysis algorithms to identify and extract individual characters from the segmented strokes.

Ligature Handling

- Ligature Recognition: Develop algorithms that can accurately detect and segment ligatures into their constituent characters. This can involve the use of pattern recognition techniques, rule-based methods, or machine learning models.

- Ligature Database: Create a comprehensive database of ligatures with corresponding constituent characters to aid in accurate recognition and segmentation.

Feature Extraction

- Structural Features: Extract structural features such as stroke direction, stroke length, and curvature to capture the unique characteristics of handwritten Bangla script.

- Statistical Features: Compute statistical features like histograms, gradient profiles, or shape context descriptors to represent the shape and texture information of characters.

Machine Learning and Deep Learning Approaches

- Supervised Learning: Train OCR models using annotated datasets of handwritten Bangla documents to learn the patterns and variations in character shapes, strokes, and ligatures.

- Convolutional Neural Networks (CNNs): Utilize CNN architectures tailored to handwritten Bangla script to perform character recognition, leveraging their ability to learn hierarchical features.

- Recurrent Neural Networks (RNNs): Employ RNNs, such as Long Short-Term Memory (LSTM) networks, to handle the sequential nature of handwritten Bangla script and capture contextual dependencies.

Dataset Augmentation and Synthesis

- Augmentation: Expand the available training data by applying data augmentation techniques, such as rotation, scaling, translation, and adding noise, to increase the diversity of samples.

- Synthesis: Generate synthetic handwritten Bangla datasets using generative models or rule-based approaches, ensuring a broader range of variations in character shapes, ligatures, and writing styles.

By combining these proposed solutions and methodologies, OCR systems can be designed specifically for handwritten Bangla documents. These approaches aim to enhance character recognition accuracy, improve segmentation and ligature handling, and mitigate the impact of noise and irregularities, ultimately leading to more reliable and efficient OCR of handwritten Bangla script.

V.CONCLUSION

In conclusion, OCR of handwritten Bangla documents presents unique challenges due to the complex characteristics of the script and variations in writing styles. However, with the advancements in OCR technologies and the availability of specialized tools, significant progress has been made in addressing these challenges. OCR systems can be optimized to effectively recognize and extract text from handwritten Bangla documents, enabling digitization, searchability and accessibility of this valuable content. The advancements in OCR technology will contribute to preserving and leveraging the vast amount of handwritten Bangla documents for research, education, and cultural preservation purposes. Further research and development efforts are required to refine the existing approaches, explore new techniques, expand annotated datasets, and adapt OCR systems to the evolving needs of handwritten Bangla script recognition.

REFERENCES: -

1. Chaudhury, Ayan & Mukherjee, Partha & Das, Sudip & Biswas, Chandan & Bhattacharya, Ujjwal. (2022). A Deep OCR for Degraded Bangla Documents. ACM Transactions on Asian and Low-Resource Language Information Processing. 21. 10.1145/3511807.
2. Md. Abir, Abu & Rahman, Sanjana & Ellin, Samia & Farzana, Maisha & Manik, Md & Rahman, Rafeed. (2020). Constraints in Developing a Complete Bengali Optical Character Recognition System.
3. Rahul Pramanik and Soumen Bag. Segmentation-based recognition system for handwritten bangla and devanagari words using conventional classification and transfer learning. IET Image Processing, 14(5):959– 972, 2020.
4. 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.
5. Rakshit, Payel & Halder, Chayan & Roy, Kaushik. (2019). An Approach toward Character Recognition of Bangla Handwritten Isolated Characters. 10.1201/9780429277573-2.
6. 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.
7. Rahul Pramanik and Soumen Bag. Shape decomposition-based handwritten compound character recognition for bangla ocr. Journal of Visual Communication and Image Representation, 50:123–134, 2018.
8. Payel Rakshit, Chayan Halder, Subhankar Ghosh, and Kaushik Roy. Line, word, and character segmentation from bangla handwritten text—a precursor toward bangla hocr. In Advanced Computing and Systems for Security, pages 109–120. Springer, 2018.

9. N. Das, S. Basu, R. Sarkar, M. Kundu, M. Nasipuri and D.K. Basu, "Handwritten Bangla Compound Character Recognition: Potential Challenges and Probable Solution", Proceedings of the International Conference on Artificial Intelligence, pp. 1901-1913, 2009.