Versatile Bengali OCR: Document Analysis Technique for Varied Document Styles and Content

AKM Shahariar Azad Rabby[1], Hasmot Ali[1], Md. Majedul Islam[1], Fuad Rahman[2]
[1]*Apurba Technologies Ltd, Dhaka, Bangladesh*
[2]*Apurba Technologies, CA, USA*
{*rabby, majed, fuad*}*@apurbatech.com, hasmot_ali@apurba.com.bd*

Abstract—**In our research paper, we introduce a distinctive Bengali OCR system that boasts impressive capabilities. This system excels in reconstructing document layouts while maintaining the integrity of structure, alignment, and even images. It integrates advanced image and signature detection for precise extraction. Specifically, tailored models for word segmentation accommodate various document types, such as computer-compose, letterpress, typewritten, and handwritten documents. Notably, the system handles static and dynamic handwritten inputs, recognizing diverse writing styles. Additionally, it achieves remarkable recognition of compound characters in the Bengali language. The comprehensive data collection contributes to a diverse corpus, and sophisticated technical components enhance character and word recognition. Other notable features include image, logo, signature recognition, table recognition, perspective correction, layout reconstruction, and a queuing module for efficient and scalable processing. The system showcases exceptional performance in the efficient and accurate extraction and analysis of text.**

1. Introduction

The advancement of Optical Character Recognition (OCR) systems has revolutionized the digitization and analysis of textual content. This paper presents a Bengali OCR system, which exhibits exceptional capabilities and stands out among existing solutions. Our system is designed to accurately reconstruct document layouts while preserving the original structure and alignment of paragraphs, tables, and numbered lists. Additionally, it goes beyond layout reconstruction by restoring embedded images, ensuring a faithful representation of the original content. Another remarkable feature is the incorporation of advanced image and signature detection, enabling accurate identification and extraction of these elements from diverse document types.

In our research paper, we introduce an all-encompassing Bengali Optical Character Recognition (OCR) system designed to tackle the complexities associated with diverse document processing. Our system integrates dedicated models for word segmentation, tailoring its capabilities to distinct document types such as computer-composed, letterpress, typewritten, and handwritten documents. Demonstrating adaptability, it adeptly processes dynamic handwritten inputs and showcases proficiency in recognizing compound characters inherent to the Bengali language, ensuring precise and accurate character recognition.

The technical elements of our OCR system are designed to optimize both accuracy and efficiency. Utilizing sophisticated methods like automatic perspective correction and character segmentation significantly enhances recognition performance. Our Neural Network architecture, based on a self-attentional VGG model with multiple heads, leverages self-attention mechanisms to capture intricate dependencies within characters. Intelligent post-processing techniques are applied to seamlessly merge recognized characters into words, optimizing the formation of words and thereby improving overall text readability.

To enhance word recognition, we generated a synthetic dataset featuring diverse font sizes, families, and backgrounds. Real-world fine-tuning and model quantization techniques were implemented to improve performance and optimize resource utilization. Our rule-based OCR layout module successfully reconstructs document layouts with precision, effectively distinguishing between paragraph and table regions. It accurately replicates numbered list items, restores images, and excels in reconstructing tables, preserving alignment and structure.

Providing a holistic approach to text extraction and analysis, our Bengali OCR system stands out as a comprehensive solution. Through rigorous data collection, specialized models, advanced techniques, and streamlined processing, our system attains notable levels of accuracy and efficiency. The utilization of optimized technical components and methodologies ensures resource efficiency, compatibility across multiple platforms, and improved scalability. These advancements collectively establish our Bengali OCR system as a robust and innovative solution for diverse text extraction and analysis tasks.

The main contributions of this paper are as follows:

- Presented a novel Bengali OCR system that is capable of accurately recognizing text from a wide range of documents, including computer-composed, letterpress, typewriter, and handwritten documents.
- Developed specialized word and character segmentation models tailored to different document types.
- Proposed a novel layout reconstruction module that

is capable of accurately reconstructing the original structure and layout of documents.

- Implemented a queuing module that facilitates an asynchronous pipeline, improving the efficiency and scalability of the system.

2. Literature Review

Recent years have witnessed significant progress in Optical Character Recognition (OCR) technology, driven by advancements in deep learning, neural network algorithms, sophisticated data preprocessing techniques, improved layout analysis approaches, and extended multi-language support. These strides have resulted in the widespread adoption of OCR for tasks such as document analysis, data digitization, and information extraction. Tesseract by Smith [1] and OCRopus by Breuel [2] are notable OCR engines widely utilized for English OCR tasks. However, the development of state-of-the-art OCR technology for low-resource languages remains an ongoing effort, with examples such as OCR for Chinese by Li et al. [3] and OCR for Arabic by Cheung et al. [4].

Noteworthy efforts in the realm of Bangla OCR have been undertaken by various researchers. Ahmed et al. [5] proposed an effective Bangla OCR system, while Hasan et al. [6] introduced a smart OCR approach capable of recognizing both Bangla and English languages. A significant challenge in Bangla OCR lies in covering various document types, addressed by Islam et al. [7] who focused on character recognition in three types of Bangla documents: computer composed, letterpress, and typewritten text. Layout understanding, a vital aspect of advanced OCR systems, was addressed by Qiao et al. [8] in the context of English OCR with multi-modal document understanding. Mindee [9] presented an OCR system with advanced layout analysis capabilities, and Jhu et al. [10] tailored a multi-stage OCR approach for documents with intricate layouts, such as newspapers.

OCR systems have also been developed to recognize tables, logos, signatures, and images within documents. Rausch et al. [11] contributed to parsing hierarchical document structures, while Chi et al. [12] proposed an approach for extracting information from documents with complex table structures. These advancements in recognizing and processing various elements within documents have significantly expanded the capabilities of OCR technology.

This paper introduces our OCR system, excelling in recognizing images, signatures, and logos within documents. With perspective correction capabilities and accurate table recognition and reconstruction, our proposed OCR technology represents a cutting-edge solution for extracting Bangla text and facilitating information digitization.

3. Proposed Methodology

Distinguished by its unique design and capabilities, our Bengali OCR system excels in various aspects. A key strength lies in its precise reconstruction of document layouts, maintaining the original structure and alignment of paragraphs, tables, and numbered lists. This extends to the restoration of embedded images, ensuring a faithful representation of the original content. surpasses the handling of static handwritten documents, showcasing versatility in processing dynamic handwritten inputs and recognizing a broad spectrum of writing styles. Finally, it achieves exceptional results in recognizing compound characters within the Bengali language, ensuring comprehensive and precise character recognition.

3.1. Data Collection

To develop our Bengali OCR system, an extensive data collection process was undertaken, resulting in the creation of the most substantial image corpus specifically tailored for Bengali OCR development. This corpus encompasses a diverse array of data types, including computer-composed, typewriter, and letterpress documents, as well as offline and online handwritten Bengali words and characters.

The handwritten data was meticulously gathered from different regions across Bangladesh, with careful attention paid to achieving a balanced representation in terms of gender, age, occupation, and geographical factors. We have tried to include every possible handwriting variety to ensure the utmost quality of the dataset. For the creation of a gold-standard dataset, each data point in our corpus went through human annotation in a structured and controlled environment, involving three annotators and a supervisor. Additionally, we have collected a huge amount of dynamic handwritten word data, totaling 1 million words, which was written using a Wacom Graphics Tablet and included precise coordinate values for enhanced accuracy and better usability.

At the time of formulating our data collection methodology, we carefully considered multiple parameters to guarantee a balanced and traceable annotated dataset. We divided the dataset into two primary groups: Printed documents and Handwritten documents, each with its distinct sub-parameters. For printed documents, we classified them into three types: computer-composed, letterpress, and typewriter documents. Each type underwent further division based on domain (subject field), the time period of text production, and the medium of text publication.

The gathered data underwent annotation using our pre-build annotation application and tool. We assigned a label to each character and diacritic, enabling the model to utilize this mapping as a label during training. The annotation process encompassed a detailed description and analysis of the dataset, including metadata and tagging. To understand the dataset's characteristics, we have performed statistical analysis on the frequency of characters, words, compound characters, graphemes, vowel diacritics, and consonant diacritics in the input documents.

After completing the annotation process, we analyzed the annotated data to calculate the inter-annotator agreement for evaluating the annotation quality. The analysis revealed agreement percentages for computer-composed, letterpress, and typewriter characters, which were 91.8%, 80.3%, and

64.2%, respectively. Additionally, we examined the presence and absence of various graphemes, characters, vowel diacritics, and consonant diacritics in the agreed-upon data. Furthermore, scrutiny of annotators' profiles indicated that 81.3% of the data showed agreement among annotators, 15.3% exhibited disagreement, and 3.4% were skipped. These analyses provide valuable insights into annotators' contributions and the overall quality of the annotated data. The final dataset comprises 178,158 computer-composed, 51,867 typewriter, 84,295 letterpress, and 57,319 handwriting images.

3.2. Model Description

To attain precise and efficient character and word recognition, our OCR system integrates advanced techniques and methodologies. The character module comprises multiple components strategically designed to enhance recognition performance. The process commences with automatic perspective correction, mitigating distortions in document images and thereby improving image quality for subsequent processing. An integral element of character recognition is the character segmentation model, which isolates individual characters from words. This approach ensures precise recognition and diminishes the likelihood of misclassifications. Through the segmentation of characters, our system can concentrate on accurately recognizing each component, resulting in elevated overall recognition accuracy.

To enhance character recognition, we introduce an innovative self-attentional VGG-based multi-headed Neural Network architecture. This design utilizes self-attention mechanisms to capture intricate dependencies and contextual information within characters. The VGG-based backbone imparts robust feature extraction capabilities, empowering the model to acquire discriminative representations of characters. Following character recognition, the system adeptly combines the recognized characters into words using intelligent post-processing techniques. This step eliminates unnecessary spaces, ensuring that the output words closely mirror the original word structures. Through this post-processing step, the overall readability and coherence of the recognized text are significantly improved.

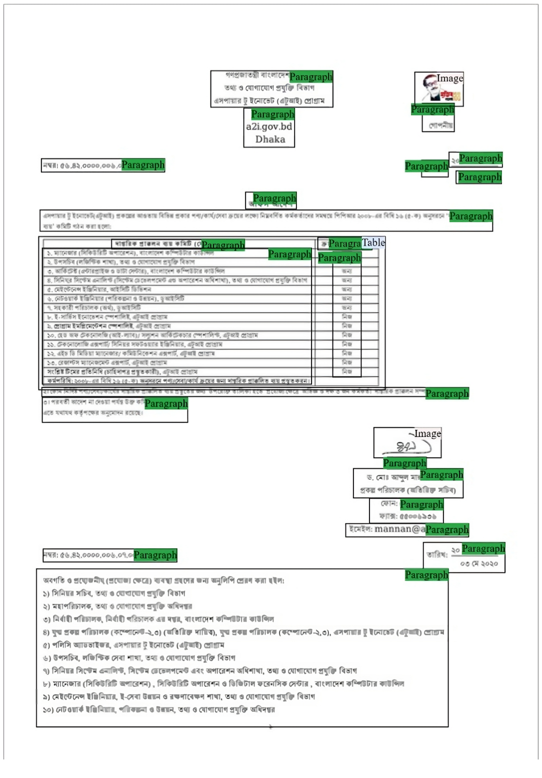
The synergy of advanced model architectures, data augmentation, fine-tuning, model quantization, and optimization for CPU deployment enables our OCR system to achieve high-performance character and word recognition across diverse document types. The incorporation of rule-based techniques complements the machine learning algorithms, ensuring robust and efficient detection.

3.3. Layout Detection and Reconstruction

Layout detection and reconstruction play a vital role in our OCR system, and we have integrated a resilient rule-based system for precise identification and reconstruction of diverse layout elements. The system adeptly discerns between paragraph and table regions, facilitating the reconstruction of lines and preserving text alignment within paragraphs, thereby maintaining the integrity of the original layout structure. Going beyond paragraph reconstruction, our layout module excels at accurately reproducing numbered list items, ensuring the faithful preservation of the structure and formatting of numbered lists. Furthermore, the module is equipped to restore images within the document, preserving their original placement and ensuring a visually accurate representation.

A significant challenge in layout reconstruction involves accurately restoring tables, and our OCR layout module excels in this aspect. It achieves flawless table reconstruction by precisely aligning the text within each cell, thereby preserving the integrity of the table structure. This capability proves particularly valuable in scenarios where tables are essential for data representation and analysis. By integrating this rule-based system into our OCR layout module, we elevate the overall accuracy and faithfulness of the reconstructed document layout. The module's proficiency in accurately reconstructing paragraphs, lists, tables, and images offers a comprehensive solution for capturing the original structure and layout of the processed documents. This ensures that the output documents maintain the same visual representation as the originals, enabling users to interact with the content in a familiar and intuitive manner.

Fig. 1 shows the segmented layout structure.



গণপ্রজাতন্ত্রী বাংলাদেশ
তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
এসপায়ার টু ইনোভেট (এটুআই) প্রোগ্রাম
a2i.gov.bd
Dhaka

গোপনীয়

নম্বর: ৫৬.৪২.০০০০.০০৬.০

২০

এসপায়ার টু ইনোভেট(এটুআই) প্রকল্পের আওতায় বিভিন্ন প্রকার পণ্য/কার্য/সেবা ক্রয়ের লক্ষ্যে নিম্নবর্ণিত কর্মকর্তাদের সমন্বয়ে পিপিআর ২০০৮-এর বিধি ১৬ (৫-ক) অনুসরণে " ব্যয়' কমিটি গঠন করা হলো:

দাপ্তরিক প্রাক্কলন ব্যয় কমিটি (	ক
১. ম্যানেজার (সিকিউরিটি অপারেশন), বাংলাদেশ কম্পিউটার কাউন্সিল	
২. উপসচিব (লজিস্টিক শাখা), তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ	
৩. আর্কিটেক্ট (এন্টারপ্রাইজ ও ডাটা সেন্টার), বাংলাদেশ কম্পিউটার কাউন্সিল	অন্য
৪. সিনিয়র সিস্টেম এনালিস্ট (সিস্টেম ডেভেলপমেন্ট এন্ড অপারেশন অধিশাখা), তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ	অন্য
৫. মেইন্টেনেন্স ইঞ্জিনিয়ার, আইসিটি ডিভিশন	অন্য
৬. নেটওয়ার্ক ইঞ্জিনিয়ার (পরিকল্পনা ও উন্নয়ন), ডুআইসিটি	অন্য
৭. সহকারী পরিচালক (অর্থ), ডুআইসিটি	অন্য
৮. ই-লার্নিং ইনোভেশন স্পেশালিস্ট, এটুআই প্রোগ্রাম	নিজ
৯. প্রোগ্রাম ইমপ্লিমেন্টেশন স্পেশালিস্ট, এটুআই প্রোগ্রাম	নিজ
১০. হেড অফ টেকনোলজি (আই-ল্যাব)/ সলুশন আর্কিটেকচার স্পেশালিস্ট, এটুআই প্রোগ্রাম	নিজ
১১. টেকনোলজি এক্সপার্ট/ সিনিয়র সফটওয়্যার ইঞ্জিনিয়ার, এটুআই প্রোগ্রাম	নিজ
১২. এইচ ডি মিডিয়া ম্যানেজার/ কমিউনিকেশন এক্সপার্ট, এটুআই প্রোগ্রাম	নিজ
১৩. প্রেজেন্টস ম্যানেজমেন্ট এক্সপার্ট, এটুআই প্রোগ্রাম	নিজ
সংশ্লিষ্ট টিমের প্রতিনিধি (চাহিদাপত্র প্রস্তুতকারী), এটুআই প্রোগ্রাম	নিজ
কর্মপরিধি: ২০০৮-এর বিধি ১৬ (৫-ক) অনুসরণে পণ্য/সেবা/কার্য ক্রয়ের জন্য দাপ্তরিক প্রাক্কলিত ব্যয় প্রস্তুতকরন।	

২। কোন নির্দিষ্ট পণ্য/সেবা/কার্যের [illegible]

৩। পরবর্তী আদেশ না দেওয়া পর্যন্ত উক্ত ক
এতে যথাযথ কর্তৃপক্ষের অনুমোদন রয়েছে।

ড. মোঃ আব্দুল মা
প্রকল্প পরিচালক (অতিরিক্ত সচিব)
ফোন:
ফ্যাক্স: ৫৫০০৬১৩১
ইমেইল: mannan@a

নম্বর: ৫৬.৪২.০০০০.০০৬.০৭.০

তারিখ: ২০
০৩ মে ২০২০

অবগতি ও প্রয়োজনীয় (প্রযোজ্য ক্ষেত্রে) ব্যবস্থা গ্রহণের জন্য অনুলিপি প্রেরণ করা হইল:
১) সিনিয়র সচিব, তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
২) মহাপরিচালক, তথ্য ও যোগাযোগ প্রযুক্তি অধিদপ্তর
৩) নির্বাহী পরিচালক, নির্বাহী পরিচালক এর দপ্তর, বাংলাদেশ কম্পিউটার কাউন্সিল
৪) যুগ্ম প্রকল্প পরিচালক (কম্পোনেন্ট-২.৩) (অতিরিক্ত দায়িত্ব), যুগ্ম প্রকল্প পরিচালক (কম্পোনেন্ট-২.৩), এসপায়ার টু ইনোভেট (এটুআই) প্রোগ্রাম
৫) পলিসি অ্যাডভাইজার, এসপায়ার টু ইনোভেট (এটুআই) প্রোগ্রাম
৬) উপসচিব, লজিস্টিক সেবা শাখা, তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
৭) সিনিয়র সিস্টেম এনালিস্ট, সিস্টেম ডেভেলপমেন্ট এবং অপারেশন অধিশাখা, তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
৮) ম্যানেজার (সিকিউরিটি অপারেশন) , সিকিউরিটি অপারেশন ও ডিজিটাল ফরেনসিক সেন্টার , বাংলাদেশ কম্পিউটার কাউন্সিল
৯) মেইন্টেনেন্স ইঞ্জিনিয়ার, ই-সেবা উন্নয়ন ও রক্ষণাবেক্ষণ শাখা, তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
১০) নেটওয়ার্ক ইঞ্জিনিয়ার, পরিকল্পনা ও উন্নয়ন, তথ্য ও যোগাযোগ প্রযুক্তি অধিদপ্তর

Figure 1. Layout Analysis

3.4. Deployment and Scalability

Efficient deployment and scalability are pivotal elements of our OCR system, and we have implemented numerous strategies to optimize performance and manage substantial workloads. To begin with, we introduced a queuing module utilizing Apache Kafka and Zookeeper, establishing an asynchronous pipeline for image processing. This queuing system guarantees a seamless and uninterrupted workflow by effectively orchestrating the flow of images through diverse processing steps. Each image is queued and processed asynchronously, eliminating the necessity for synchronous requests and consequently enhancing the overall processing speed.

To ensure multi-platform compatibility, we have converted our trained models to the ONNX [13] format, which facilitates seamless integration and deployment across different deep-learning frameworks and platforms. Additionally, we have leveraged the Triton [14] framework for efficient model deployment and scaling in production environments. Triton provides a robust infrastructure that allows us to handle large workloads and ensures reliable performance. Our OCR system takes an average time of 1.847 seconds for Computer Compose, 3.319 seconds for Handwritten, 1.756 seconds for Typewriters and 3.276 seconds for Letterpress documents.

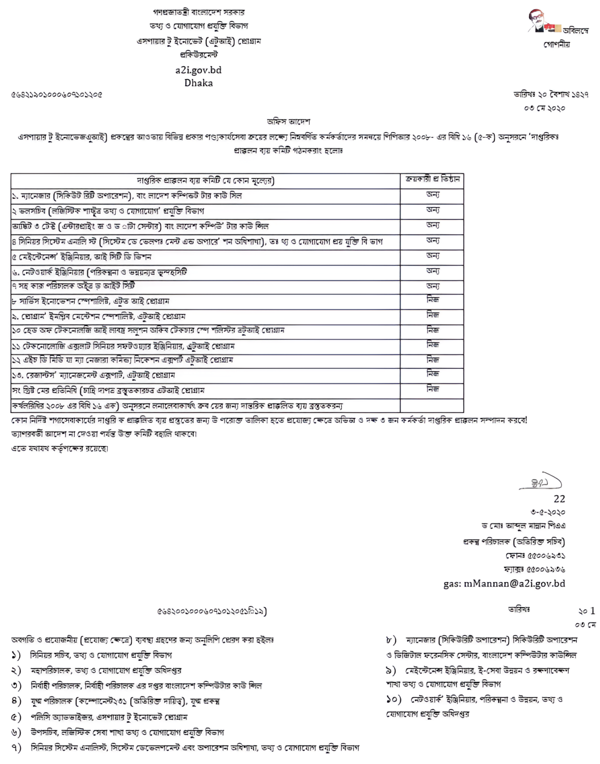
By combining these strategies, our OCR system achieves enhanced efficiency, scalability, and reliability. The queuing module, based on Apache Kafka [15] and Zookeeper [16], enables smooth image processing, reduces timeouts, and efficiently handles large files. The compatibility with ONNX format ensures easy integration with different frameworks and platforms, while the use of the Triton framework provides a scalable and robust infrastructure for deployment in production environments. These measures collectively contribute to a high-performance OCR solution capable of meeting the demands of complex OCR tasks.

4. Performance Evaluation

The performance of our OCR model was evaluated using randomly selected documents from each document type, including computer-composed (CC), handwritten (HWR), typewritten (TW), and letterpress (LP). These documents varied in font types, sizes, noise levels, and sources. Each document was processed through the inference pipeline, which involved noise removal and word boundary detection using docTR for character segmentation. The segmented characters were then passed to the model for grapheme prediction, word and sentence generation. Fig. 2 shows the original input image with the constructed output with proper layout.

The accuracy of the OCR output was measured by comparing it with the ground truth using the Levenshtein distance [17].

The accuracy results, Levenshtein distance-based, for each document type are presented in Table. 1. The Levenshtein distance-based accuracy measures the similarity between the OCR output and the ground truth.



গণপ্রজাতন্ত্রী বাংলাদেশ সরকার
তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
এসপায়ার টু ইনোভেট (এটুআই) প্রোগ্রাম
প্রকিউরমেন্ট
a2i.gov.bd
Dhaka

অবিলম্বে
গোপনীয়

৫৬৪২১১৯০১০০০৬০৭১০১২০৫

তারিখঃ ২০ বৈশাখ ১৪২৭
০৩ মে ২০২০

অফিস আদেশ

এসপায়ার টু ইনোভেজএআই) প্রকল্পের আওতায় বিভিন্ন প্রকার পণ্যকার্যসেবা ক্রয়ের লক্ষ্যে নিম্নবর্ণিত কর্মকর্তাদের সমন্বয়ে পিপিআর ২০০৮- এর বিধি ১৬ (৫-ক) অনুসরনে 'দাপ্তরিক প্রাক্কলন ব্যয় কমিটি গঠনক্যাং হলোঃ

দাপ্তরিক প্রাক্কলন ব্যয় কমিটি যে কোন মূল্যের)	ক্রয়কারী প্র তিষ্ঠান
১. ম্যানেজার (সিকিউট রিটি অপারেশন), বাং লাদেশ কম্পিউট টার কাউ সিল	অন্য
২ উপসচিব (লজিস্টিক শাখার তথ্য ও যোগাযোগ' প্রযুক্তি বিভাগ	অন্য
অডিট ও টেস্ট (এন্টারপ্রাইং জ ও ড ⌂টা সেন্টার) বাং লাদেশ কম্পিউ' টার কাউ ন্সিল	অন্য
৪ সিনিয়র সিস্টেম এনালি স্ট (সিস্টেম ডে ভেলপম মেন্ট এন্ড অপারে' শন অধিশাখা), তঃ থ্য ও যোগাযোগ প্রয় যুক্তি বি ভাগ	অন্য
৫ মেইন্টেনেন্স' ইঞ্জিনিয়ার, আই সিটি ডি ভিশন	অন্য
৬. নেটওয়ার্ক ইঞ্জিনিয়ার (পরিকল্পনা ও উন্নয়নের ভূম্যহসিটি	অন্য
৭ সহ কারী পরিচালক অইর ড আইটি সিটি	অন্য
৮ সার্ভিস ইনোভেশন স্পেশালিষ্ট, এটুত আই প্রোগ্রাম	নিজ
৯. প্রোগ্রাম' ইনগ্রিব নেন্টেশন স্পেশালিষ্ট, এটুআই প্রোগ্রাম	নিজ
১০ হেড অব টেকনোলজি আই লাবর সলুশন অফিস টেকুয়ার স্পে শলিস্টর বটুআই প্রোগ্রাম	নিজ
১১ টেকনোলোজি এক্সপার্ট সিনিয়র সফটওয়্যার ইঞ্জিনিয়ার, এটুআই প্রোগ্রাম	নিজ
১২ এইচ ডি মিডি য় ন্যা নেজারা কমিভ্য নিকেশন এক্সপার্ট এটুআই প্রোগ্রাম	নিজ
১৩. রেক্রান্টস' ম্যানেজমেন্ট এক্সপার্ট, এটুআই প্রোগ্রাম	নিজ
সং স্ট্রিষ্ট মেত্র প্রতিনিধি (চাহি দাপত্র প্রস্তুতকারক এটুআই প্রোগ্রাম	নিজ
কর্মনিপিথি ২০০৮ এর বিধি ১৬ এক) অনুসরনে বনানেবাকার্যং ক্রয় য়ের জন্য দাপ্তরিক প্রাক্কলিত ব্যয় প্রস্তুতকরন্য	

কোন নির্দিষ্ট পণ্যসেবাকার্যের দাপ্তরি ক প্রাক্কলিত ব্যয় প্রস্তুতের জন্য উ পরোক্ত তালিকা হতে প্রযোজ্য ক্ষেত্রে অভিজ্ঞ ও দক্ষ ৩ জন কর্মকর্তা দাপ্তরিক প্রাক্কলন সম্পাদন করবে!
ত্যাপরবর্তী আদেশ না দেওয়া পর্যন্ত উক্ত কমিটি বহাল থাকবে।
এতে যথাযথ কর্তৃপক্ষের রয়েছে।

22
৩-৫-২০২০
ড মোঃ আব্দুল মান্নান পিএএ
প্রকল্প পরিচালক (অতিরিক্ত সচিব)
ফোনঃ ৫৫০০৬২৩১
ফ্যাক্সঃ ৫৫০০৬১৩৬
gas: mMannan@a2i.gov.bd

৫৬৪২০০১০০০৬০৭১০১২০৫১৫১১)

তারিখঃ ২০ 1
০৩ মে

অবগতি ও প্রয়োজনীয় (প্রযোজ্য ক্ষেত্রে) ব্যবস্থা গ্রহণের জন্য অনুলিপি প্রেরণ করা হইলঃ
১) সিনিয়র সচিব, তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
২) মহাপরিচালক, তথ্য ও যোগাযোগ প্রযুক্তি অধিদপ্তর
৩) নির্বাহী পরিচালক, নির্বাহী পরিচালক এর দপ্তর বাংলাদেশ কম্পিউটার কাউ ন্সিল
৪) যুগ্ম পরিচালক (কম্পোনেন্ট২৩১ (অতিরিক্ত দায়িত্ব), যুগ্ম প্রকল্প
৫) পলিসি অ্যাডভাইজর, এসপায়ার টু ইনোভেট প্রোগ্রাম
৬) উপসচিব, লজিস্টিক সেবা শাখা তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
৭) সিনিয়র সিস্টেম এনালিস্ট, সিস্টেম ডেভেলপমেন্ট এবং অপারেশন অধিশাখা, তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
৮) ম্যানেজার (সিকিউরিটি অপারেশন) সিকিউরিটি অপারেশন ও ডিজিটাল ফরেনসিক সেন্টার, বাংলাদেশ কম্পিউটার কাউন্সিল
৯) মেইন্টেনেন্স ইঞ্জিনিয়ার, ই-সেবা উন্নয়ন ও রক্ষণাবেক্ষণ শাখা তথ্য ও যোগাযোগ প্রযুক্তি বিভাগ
১০) নেটওয়ার্ক ইঞ্জিনিয়ার, পরিকল্পনা ও উন্নয়ন, তথ্য ও যোগাযোগ প্রযুক্তি অধিদপ্তর

Figure 2. Layout Reconstructed OCR Output

TABLE 1. LEVENSHTEIN DISTANCE-BASED ACCURACY

Document Type	Accuracy
Computer compose	90.06%
Letterpress	88.53%
Typewriter	83.38%
Handwritten	86.84%
Average	87.20%

On average, our model achieved a confusion matrix-based accuracy of 98.05% and a Levenshtein distance-based accuracy of 87.20%. Comparing our results with recent state-of-the-art works on Bangla OCR, which primarily focused on handwritten data, our model demonstrates competitive performance across all document types.

5. Conclusion

In conclusion, our research paper introduces a comprehensive and innovative Bengali OCR system distinguished by its unique design and outstanding capabilities. The system excels in reconstructing document layouts, preserving the structure and alignment of paragraphs, tables, and numbered lists. Beyond this, it adeptly restores embedded images, maintaining fidelity to the original content. Advanced image and signature detection further enhance the system's accuracy and versatility. Tailored models for word segmentation accommodate various document types, and the OCR system seamlessly handles both static and dynamic handwritten inputs. Exceptional recognition of compound characters in Bengali ensures thorough character recognition. The system's technical components optimize character and word recognition through features like automatic perspective correction and self-attentional neural networks. Additionally, the integration of queuing mechanisms facilitates efficient and scalable processing, particularly for handling large files. These advancements position our Bengali OCR system as a highly effective solution for efficient and accurate text extraction and analysis in Bengali documents.

Acknowledgment

The authors would like to acknowledge the encouragement and funding from the "Enhancement of Bangla Language in ICT through Research & Development (EBLICT)" project, under the Ministry of ICT, the Government of Bangladesh.

References

[1] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.

[2] Breuel, T.M. (2008). The OCRopus open source OCR system. Electronic imaging.

[3] Qiang Huo, Yong Ge and Zhi-Dan Feng, "High performance Chinese OCR based on Gabor features, discriminative feature extraction and model training," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 2001, pp. 1517-1520 vol.3, doi: 10.1109/ICASSP.2001.941220.

[4] Cheung, A., Bennamoun, M., and Bergmann, N. W., "An Arabic optical character recognition system using recognition-based segmentation", Pattern Recognition, vol. 34, no. 2, pp. 215–233, 2001. doi:10.1016/S0031-3203(99)00227-7.

[5] T. Ahmed, M. N. Raihan, R. Kushol and M. S. Salekin, "A Complete Bangla Optical Character Recognition System: An Effective Approach," 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019, pp. 1-7, doi: 10.1109/ICCIT48885.2019.9038551.

[6] M. R. Hasan et al., "Smart OCR for Recognizing Bangla Characters with CRAFT and Deep Learning Models," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, NY, USA, 2022, pp. 0573-0577, doi: 10.1109/UEMCON54665.2022.9965668.

[7] M. M. Islam, A. Das, I. Kowsar, A. K. M. Shahariar Azad Rabby, N. Hasan and F. Rahman, "Towards building a Bangla text recognition solution with a Multi-Headed CNN architecture," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 1061-1067, doi: 10.1109/BigData52589.2021.9671653.

[8] Qiao, Liang, et al. "DavarOCR: A Toolbox for OCR and Multi-Modal Document Understanding." Proceedings of the 30th ACM International Conference on Multimedia. 2022.

[9] Mindee. 2021. doctr: Document text recognition. https://github.com/mindee/doctr.

[10] Zhu, W., Sokhandan, N., Yang, G., Martin, S., & Sathyanarayana, S. (2022). DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), 12643-12649. https://doi.org/10.1609/aaai.v36i11.21539

[11] Rausch, J., Martinez, O., Bissig, F., Zhang, C., & Feuerriegel, S. (2021). DocParser: Hierarchical Document Structure Parsing from Renderings. Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), 4328-4338. https://doi.org/10.1609/aaai.v35i5.16558

[12] Chi, Zewen, et al. "Complicated table structure recognition." arXiv preprint arXiv:1908.04729 (2019).

[13] ONNX. 2017. Open standard for machine learning interoperability. https://github.com/onnx/onnx.

[14] NVIDIA Corporation. Triton Inference Server: An Optimized Cloud and Edge Inferencing Solution. https://github.com/triton-inference-server/server

[15] Sax, M.J. (2018). Apache Kafka. In: Sakr, S., Zomaya, A. (eds) Encyclopedia of Big Data Technologies. Springer, Cham. https://doi.org/10.1007/978-3-319-63962-8_196-1

[16] Apache Software Foundation. 2011. Zookeeper. Letzter Zugriff: 11. October 2011. https://zookeeper.apache.org/doc/current/zookeeperOver.html

[17] L. Yujian and L. Bo, "A Normalized Levenshtein Distance Metric," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1091-1095, June 2007, doi: 10.1109/TPAMI.2007.1078.