Gold Standard Bangla OCR Dataset: An In-Depth Look at Data Preprocessing and Annotation Processes

Hasmot Ali[1], AKM Shahariar Azad Rabby[1,3], Md. Majedul Islam[1], A.K.M Mahamud[1], Nazmul Hasan[1], Fuad Rahman[2]
[1]Apurba Technologies, Dhaka, Bangladesh, [2]Apurba Technologies, CA, USA
[3]The University of Alabama at Birmingham, AL, USA
hasmot_ali@apurba.com.bd, {rabby, majed, fakhru, nazmul, fuad}@apurbatech.com

Abstract

This research paper focuses on developing an improved Bangla Optical Character Recognition (OCR) system, addressing the challenges posed by the complexity of Bangla text structure, diverse handwriting styles, and the scarcity of comprehensive datasets. Leveraging recent advancements in Deep Learning and OCR techniques, we anticipate a significant enhancement in the performance of Bangla OCR by utilizing a large and diverse collection of labeled Bangla text image datasets. This study introduces the most extensive gold standard corpus for Bangla characters and words, comprising over 4 million human-annotated images. Our dataset encompasses various document types, such as Computer Compose, Letterpress, Typewriters, Outdoor Banner-Poster, and Handwritten documents, gathered from diverse sources. The entire corpus has undergone meticulous human annotation, employing a controlled annotation procedure consisting of three-step annotation and one-step validation, ensuring adherence to gold standard criteria. This paper provides a comprehensive overview of the complete data collection procedure. The ICT Division, Government of the People's Republic of Bangladesh, will make the dataset publicly available, facilitating further research and development in Bangla OCR and related domains.

1 Introduction

Optical Character Recognition (OCR) technology has revolutionized the automation of data extraction from printed or handwritten text, enabling the conversion of scanned documents or image files into machine-readable formats for efficient data processing and information retrieval. While OCR applications have been extensively developed for various languages, such as English, datasets designed explicitly for the Bengali language have been limited. Existing datasets like Uber-Text(Zhang et al., 2017) have 110k images and 4.84 text instances per image on average, RoadText-1K(Reddy et al., 2020) have 1000 video clips of driving with annotations for text bounding boxes and transcriptions in every frame, TextOCR(Singh et al., 2021) have 900k annotated words collected on real images, Brno(Kišš et al., 2019) contains 19725 photographs and more than 500k text lines with precise transcriptions, and COCO-Text(Veit et al., 2016) contains over 173k text annotations in over 63k images, have primarily focused on English language OCR tasks.

In the Bengali language, datasets for OCR tasks have mainly focused on isolated character recognition (Islam et al., 2021; Das et al., 2022), with examples such as BanglaLekha-Isolated (Biswas et al., 2017), EkushNet (Rabby et al., 2018) and Borno (Rabby et al., 2021). Some authors also tried to recognize only digits (Haque et al., 2018). OCR technology requires methods, algorithms, and datasets to accurately recognize characters from continuous text with complex layouts. This is why an actual implementation of OCR in the Bengali language has yet to be presented. Some OCR technology in the Bengali language is introduced for computer-composed printed documents and character recognition on some handwritten documents. But a fully-featured Bengali OCR with character and continuous word recognition for computer compose, handwritten, and other forms of documents has yet to be presented. One of the major reasons for not presenting such an OCR application is the lack of a perfect dataset to build an efficient OCR application.

To address this gap, we present a novel and extensive dataset that includes over 4 million annotated words and characters, covering various forms of Bengali script. The primary objective of this research contribution is to provide a high-quality dataset that facilitates the advancement of Bangla OCR technology and accelerates research and development in Bengali language processing. By

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 460–470

offering a comprehensive dataset encompassing different document types and writing styles, we aim to support the creation of accurate and robust OCR systems tailored for low research languages, specifically for Bengali.

The availability of this dataset is expected to foster significant advancements in Bengali OCR research. Researchers and developers can utilize this dataset to train and evaluate OCR models, thereby enhancing the accuracy and performance of OCR systems for the Bengali language, which can lead to the improvement of technology advancements like document digitization, large-scale text analysis, linguistic studies, and historical research as well as socially impactful application like cultural heritage preservation, cross-linguistic studies, and community engagement.

2 Literature Review

Text Recognition has gained significant importance in recent years, particularly for extracting information from existing written documents. While previous methods relied on manual composition, the development of OCR technology has provided automated solutions, improving accuracy and efficiency in data extraction.

Various OCR systems have been developed for formal text in the context of the English language OCR, but recognizing handwritten material remains challenging. The IAM-database (Marti and Bunke, 2002) is a collection of English Sentence corpus containing texts that comprehend about one million word instances. They also include some image-processing procedures for extracting and segmenting the handwritten text into lines and words. For offline handwriting recognition over the International Arabic Recognition competition dataset, (Graves and Schmidhuber, 2008) represents a multidimensional recurrent neural network with the combination of connectionist temporal classification, which outperformed all entries with an accuracy of 91.4%.

A dataset of 4,587 Arabic articles with 8,994 images is presented by Doush et al. (2018) for Arabic printed documents. Firmani et al. (2017) focused on constructing an OCR system for Latin letters. From register 12 of Pope Honorii III, a proprietary crowdsourcing platform employed high-resolution (300 dpi, 2136 × 2697 pixels) scans of 30 pages to annotate a corpus of Latin letters.

For the Urdu language, Ahmed et al. (2019) introduced the Urdu-Nasta'liq Handwritten Dataset, which comprises natural handwriting from 500 writers on A4 size paper. The dataset includes commonly used ligatures and has been made publicly available. The researchers employed recurrent neural networks and achieved high accuracy in handwritten Urdu character recognition. Mainkar et al. (2020) developed a system that identifies and converts handwritten data into editable text. Their system achieved 90% accuracy in recognizing handwritten papers, providing a convenient way to modify or distribute the captured data.

In terms of data collection methods, the use of mobile applications has gained popularity due to their ease and efficiency. Azad Rabby et al. (2018) showed a universal way to collect and process handwritten data from any language. Using that method, Rabby et al. (2019) and Ferdous et al. (2021) created two datasets containing 673,482 characters. Robby et al. (2019) collected a dataset containing 5880 characters for OCR in non-Latin characters, specifically Japanese characters, using mobile apps. They trained various models with the collected dataset and Tesseract OCR tools.

For the Bangla language, BanglaLekha-Isolated (Biswas et al., 2017) has a collection of 166,105 handwritten character images having 84 different characters comprising 50 Bangla basic characters, 10 Bangla numerals, and 24 selected compound characters. Ekush (Rabby et al., 2019) is a collection of Bangla modifiers, vowels, consonants, compound letters, and numerical digits summing 367,018 isolated handwritten characters written by 3086 unique writers. Mentioned two of the datasets are for handwritten documents and the collection of character-level data. A character level but for printed documents image corpus is presented by Rifat et al. (2021) where they presented a collection of Machine Annotated eight and a half million Bangla characters. A word label synthetic printed dataset is presented by Roy et al. (2023), which contains 2 million sample images varied in fonts, domain, and vocabulary size. Also, Shihab et al. (2023), a large-scale document layout analysis dataset having 33,695 human-annotated document samples for Bangla documents, will enrich the performance of the Bangla OCR system.

The reviewed literature demonstrates the advancements and challenges in OCR technology for various languages, including English and Urdu. However, more research needs to be conducted on

Bengali OCR systems. The research presented in this paper aims to address this gap by providing a comprehensive dataset and proposing innovative methodologies for Bengali OCR. The following section will present the methodology employed in the data collection and annotation process, followed by a detailed dataset analysis in subsequent sections.

3 Data Collection Methodology

The dataset comprises various types of Bengali written documents, segmented into character and word images. Our objective is to create a comprehensive dataset for Bangla written documents, catering to the application of OCR and facilitating research in related fields. Therefore, we aimed to cover all aspects of data sources used in Bangla document writing.

3.1 Target Documents

Our data collection targets all types of Bangla documents used for written communication. These include computer-composed characters, computer-composed isolated words, computer-composed running words, letterpress-composed characters, letterpress-composed isolated words, letterpress-composed running words, typewriter-composed characters, typewriter-composed isolated words, typewriter-composed running words, handwritten characters, handwritten isolated words, handwritten running words, dynamic handwriting, and outdoor-captured data.

3.2 Data Collection Challenges

As the first major Bangla OCR project, we faced the challenge of needing a prior plan, template, or guidelines to follow. We had to iteratively plan and execute our approach until we achieved the desired results.

Technology Challenge: Collecting handwritten Bangla OCR data required manual collection from various groups of people. We scanned each handwritten document page to convert it into image format using a scanner. To ensure the highest quality of handwriting data, we generated images from documents with a specific dimension of 4938x6992 pixels and a resolution of 600 dpi while scanning. We didn't perform any image processing tasks in this step. Freshly scanned images are sent for further processing. Outdoor data was collected using a mobile or digital camera with at least a 16-megapixel camera setup. We used the "Wacom One By CTL-672/K2-F Medium Dimensions 18.9 x 27.7 x 0.9 Cm Pen Graphics Tablet" to collect dynamic handwriting data.

Resource Challenge: The computer-composed, letterpress, and typewritten data were collected from various books, newspapers, documents, and articles. Collecting letterpress and typewriter documents posed a challenge as they are less commonly used nowadays and, thus, more challenging to find. Handwritten data were collected from individuals, and finding a diverse group of writers in terms of age, educational background, occupation, etc., was also challenging. We have formed a group of more than 150 Data Processing Engineers (DPE) and five Research Assistants (RA) for this data collection procedure.

Software Challenge: Since our work is the first major Bangla OCR project, we did not have prebuilt software for data collection and preprocessing. We had to develop our tools for these purposes.

Validation Challenge: Validating our work required comparison with previous works. However, we did not find any prior work that closely matched ours, which presented challenges in the validation process.

The impact of the COVID-19 pandemic exacerbated all the aforementioned challenges. Performing tasks according to our planned timeline during the pandemic was incredibly challenging. We faced difficulties coordinating resources and challenges in contracting with writers, people responsible for scanning documents, and data collectors.

3.3 Data Source

All data sources, except handwritten documents, outdoor images, and dynamic handwriting, were provided by the ICT Division under the Ministry of Posts, Telecommunications, and Information Technology, Government of Bangladesh. Outdoor images were physically collected from different locations in Dhaka, Narayangonj, Rajshahi, and Barisal districts. The data collection for outdoor images focused on nameplates, banners, festoons, and posters. Dynamic handwriting data was collected using a computer and a graphics tablet. The specific sources of data are mentioned below.

Computer Composed: The sources for computer-composed data include Bangla Academy Bibortonmulok Bangla Ovidhan 2, Anandamoth by Bankim Chandra Chattopadhyay, Chinmoy Bongo,

ICT Ministry Notices, Education Board, Mamlar Ejhar, Parliament, A2i, Primary Engineering Drawing, Civil Construction, Chukti Ainer Vassho, Shahitto Path, and Kaler Shongkhipto Itihash. We aimed to include all types of words in our dataset, resulting in diverse data sources. The mentioned sources include a Bangla dictionary, literature, government notices, judicial documents, parliamentary documents, engineering books, and fiction.

Letterpress: The letterpress data sources include Sonet Ponchashot o Onnano Kobita, Ekti Ful Ke Bachabo Bole, Prasuti Tantra, Desh Magazine, Chiranjib-Banoushadhi, Na Tarashonkor Bondopaddhoy, Banglar Boma, Oporadh Biggan, Shompotti Hostantor Ain, Rasta Nirmal Part: 1, 2, 3, Dhaka Nagorir Mosjid Nirdeshika, Futonto Golap, Goutam Buddho, Humayun Kabir Rochonaboli, Medica Homeopathy, Motor Shikkha, Ajad Hindi Fouz, Shompotti Hostantor Ain, Oporadh Biggan, and Dhaka Nagorir Mosjid Nirdeshika. As letterpress printed documents are relatively rare, we collected as many documents as possible. The letterpress printed documents also exhibit a wide range of word varieties.

Typewriter: The typewriter data sources include the Ministry of Health, Education Board, Water Development Board, National Parliament, Rajuk, National Archive, and Press Institute Bangladesh.

Handwritten: All handwritten data was collected from individuals residing in different districts. The process involved preparing, printing, and distributing the text script among the writers, along with blank paper and black pens. The text script contained Bangla Stories, Poem, and News collection. When collecting the script, we focused on collecting diverse word ranges, including most usages and rarely usages words among people. Every individual writer is provided with a different text script to write. After completing the writing, we collected the written pages with proper labeling to trace the text script and aid in data annotation. We then sent the collected documents for scanning to generate high-resolution images used for word cropping and annotation. We had a diverse group of writers regarding age, gender, occupation, and educational background.

3.4 Software

We employed various software tools to expedite the segmentation and annotation processes. Our team developed all the tools. We developed our cropping tools for word segmentation from the pages and annotation tools for associating the correct text with the cropped images. The annotation tools provided the opportunity to remove incorrect words and fix incorrect annotations.

Data Segmentation: Segmentation involves cropping words from the entire image. After scanning, we used our cropping tools to segment the words from the images. All handwritten word and character data were prepared using the cropping tools. Contour-based cropping is performed for computer-composed letterpress and typewriter documents.

Manual cropping was performed using our custom cropping tools. These tools were used specifically for handwritten isolated and running words and outdoor images. The interface of the tools consisted of three windows: one for displaying the scanned page image, another for displaying the script (txt file) corresponding to the page, and the last one for displaying the cropped images along with their associated scripts. Users were required to crop each word from the entire page and tag the cropped word or character image with the corresponding script. The users could view the status of the cropped images, such as what was cropped and how many data images were cropped from the page. Figure 1 shows the interface of the cropping tools.

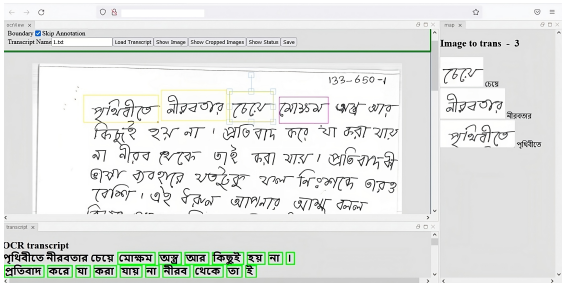

Fig. 1. Cropping Tools User Interface

Contour-based cropping was performed using a pixel-wise semantic segmentation approach. We utilized contour-based cropping for some computer-composed words, computer-composed characters, typewriter words, typewriter characters, letterpress words, and letterpress characters.

Data Annotation: After collecting segmented word images using the above cropping methods, we annotated the images with the corresponding words from the transcript. We used two different tools for word annotation; different groups of annotators and

supervisors performed and validated the annotation procedure.

Initially, a mobile application was used for word and character annotation. It had an interface consisting of an image view displaying the image to be annotated, a text field displaying the corresponding word or character for validation (matching the correct image with text), and three buttons (Reject, Skip, Accept) arranged horizontally. Users could choose from the three options to annotate the correct script or text with the corresponding image. Users could also swipe the screen right to accept, left to reject, or down to skip. If the user believed that the text/script perfectly matched the character in the given image, they would select Accept. If the image were unclear for annotation, the user would select Skip. If the user found that the text/script did not match the character in the given image but was clear enough to identify as a different character, the user would select Reject. The decision-making process for gold standard data involved accepting every single data at least 15 times out of 25 attempts by five individual annotators. Figure 2 shows the mobile app's user interface for data annotation.

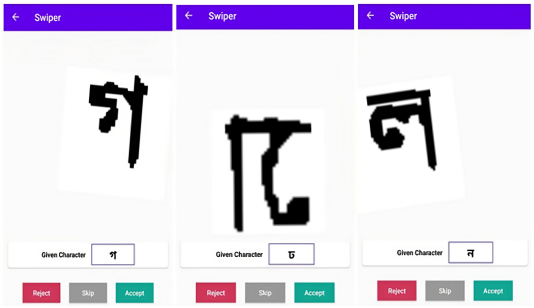

Fig. 2. Android Application User Interface.

We developed a comprehensive tool that facilitated data management tasks, including image-corpus collection, processing, annotation, indexing, and distribution. This tool served as a hub for all image data-related operations. It has image processing features such as systematically importing and storing scanned images, cropping, resizing, zooming, rotating, vectorizing, and skew editing imported image files and applying and storing processed image metadata.

The annotation web tool had an interface with word or character images associated with a text field for annotating the correct text with the image. Annotators are responsible for writing the correct word in the text field. If a word was not easily understandable or the annotator noticed that the image was cropped abnormally, they could mark the corresponding tick and clear the text box to save/drop the data.

As we collected the handwritten data with the corresponding script, the text box provided suggestions for characters or words for handwritten isolated and running words. The annotation tools functioned differently for normal annotators and supervisors. Normal annotators had the usual interface for annotating data, and there were three groups of normal annotators for annotating each data set. Figure 3 shows the user interface of the annotation tools for normal annotators.

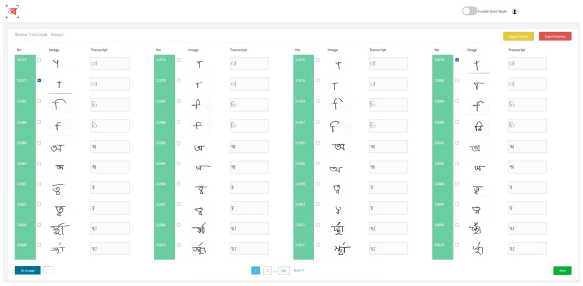

Fig. 3. Annotation Tools Interface for Normal Annotator.

Supervisors were responsible for validating the data annotated by the normal annotators. The supervisor interface had three additional columns displaying the annotated words from the three individual normal annotators. The supervisor had to choose the correct annotation among them, validating the data and finalizing the annotation for the gold standard dataset. Figure 4 shows the user interface of the annotation tools for supervisor-level annotation.

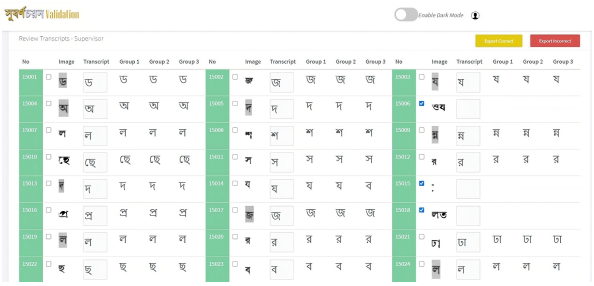

Fig. 4. Annotation Tools Interface for Supervisor-level Annotator.

3.5 Validation

We ensured the quality of the collected data through various validation phases. Data validation was a

significant concern with segmentation and annotation steps at every data collection stage.

Data Collection: During the data collection phase, we validated the quality of the handwriting. We did not solely focus on clear and beautiful handwriting; we also included samples of unclear handwriting. We ensured diversity in the most frequently used words and aimed to generate the highest-quality images during the scanning process. We validated that the scanned images were aligned correctly and not overly rotated, as excessively rotated images were excluded from cropping.

Data Segmentation: During the segmentation step, we validated that the cropped words were accurately segmented. Ensured that noise and unnecessary surroundings were not included when cropping a word. Then, carefully tagged the corresponding text script with the cropped word and cross-checked the cropped words with their corresponding text scripts before starting the annotation process.

Annotation: Initially, each data point was annotated by three individual annotators using our annotation tools. After the normal annotators completed their annotation, the data was transferred to the supervisor annotation phase, where a supervisor validated the decisions made by the three individual annotators and saved the final annotations. This resulted in the final annotation results.

Gold Standard Dataset: After validation by the three normal annotators and one supervisor annotator, the final data was stored as the gold standard dataset.

3.6 Data Statistics

The initial and essential step in building large-scale Bengali OCR models is to curate a comprehensive dataset comprising text images from various contexts and environments. The objective is to capture all possible variations in Bengali text and enhance the robustness of the AI models. Our dataset consists of 4,116,073 annotated images, which were 4,548,665 before the annotation process. The data is well-balanced regarding font type, size, noise, and data source. It is also balanced in terms of the number of words and characters. Some of the most frequently occurring words in the dataset include না, করে, এই, আমি, আর, আমার, তার, থেকে, ও, হয়, এবং while the most common characters are া, ে, ব, ি, ন, র, ক, ত, প, হ. See Appendix A for the data statistics before and after annotation with detailed data distribution statistics for each category. Every category is balanced based on specific criteria such as font type, size, noise, source, age, education, place, light, and color for specific data types.

The data statistics highlight the significant effort and careful curation that went into creating a diverse and balanced dataset, which is crucial for training accurate and robust Bengali OCR models.

3.7 Quality Assessment

To ensure the accuracy and reliability of our work, we conducted a comprehensive quality assessment of the tasks performed throughout the project. We divided the evaluation into two key components: Data Cropping and Data Annotation. Several factors were evaluated for each component, as described below.

Segmentation Quality: Since we worked with handwritten images, accurate segmentation was crucial. We implemented both automatic and manual approaches to ensure the highest quality segmentation.

Cropping Quality: We formed multiple groups of data annotators and established a hierarchical organization. Initially, we assigned 30 individuals for the cropping task. Additionally, a supervision team consisting of 5 experienced members was formed to oversee the process and maintain quality. Finally, we thoroughly inspected the pages to ensure superior quality and minimize human errors.

Annotation Quality: To achieve a gold standard dataset, we followed a meticulous process involving four individuals. Three annotators annotated words with their corresponding images, while one supervisor validated and ensured the dataset's gold standard status. After multiple rounds of careful scrutiny, we finalized the dataset based on the annotators' choices. Figure 5 visualizes the number of data before and after annotation, highlighting the annotation process's impact.

Moreover, the statistical efficiency of a dataset can be measured using the Kappa Score (McHugh, 2012). Fleiss' Kappa or Coheren's Kappa coefficient assesses the inter-annotator reliability or agreement among two or more annotators introduced by Fleiss (1971). Kappa score is calculated by Equation 1.

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

where P_o is the observed agreement among the annotators. It is the proportion of cases where

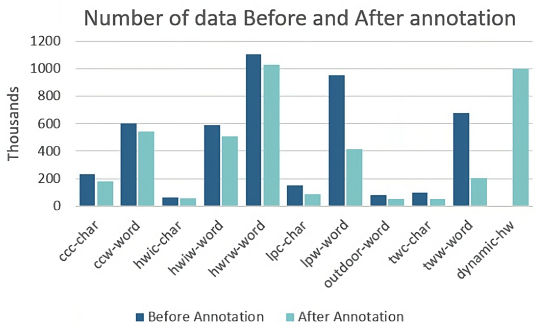

Fig. 5. number of data before and after annotation.

all three annotators agree, and P_e is the expected agreement by chance. It is calculated as the sum of the squared proportions of each category, assuming independence between the annotators.

A higher Kappa Score indicates a higher level of agreement between annotators, thereby increasing the reliability of the data validation process. A dataset is considered reliable if it achieves a satisfactory Kappa Score. In our case, we obtained a Kappa Score of 0.91 for computer-composed documents, 0.93 for Letterpress documents, and 0.78 for Typewriter documents. These scores demonstrate the high level of agreement and reliability among different-level annotators.

4 Experimental Evaluation

The presented dataset is a part of the OCR project under The ICT Division, Government of the People's Republic of Bangladesh. Due to the confidentiality of the assignment, the dataset has minimal access to use for building or testing existing models. However, one of a small subset of this data is provided to a Research and Development lab for evaluating the dataset. They have tested their CRNN-VDS model with VDS Character Representation (Roy et al., 2023) using our Computer Compose, Letterpress, and Typewriters data. CRNN-VDS is trained on a large-scale synthetic dataset having 2 million samples. Table 1 represents their experimental performance report of the CRNN-VDS model for Character Recognition Rate (CRR) and Word Recognition Rate (WRR) with this dataset.

Table 1: CRNN-VDS model performance on this dataset

Document Type	**CRR**	**WRR**
Computer compose	93.04%	79.03%
Letterpress	83.61%	57.86%
Typewriter	70.60%	28.05%

5 Discussion and Conclusion

In this research paper, we have presented the development of a monumental Bangla OCR corpus, encompassing various document types, including computer-compose, typewriter, letterpress, handwriting, and outdoor scene text. The data collection process involved meticulous steps, including document collection, scanning, cropping, annotation, and validation. Throughout the project, we focused on maintaining the highest quality standards for the collected data. However, our efficient team successfully managed this challenging process by implementing well-organized instructions and thorough supervision, ensuring minimal chances of errors.

We encountered some challenges in the segmentation and annotation phases. Auto-cropping segmentation faced difficulties with noisy documents and handwriting quality, as the quality of the paper significantly affected the legibility of the handwriting. Setting a static threshold value was also challenging, considering the variation in handwriting dimensions and stroke sizes. Consequently, we switched to manual cropping to ensure accurate results. In the annotation process, we faced complications, particularly with our initial Android application. However, we learned from this experience and later assembled teams with efficient and experienced supervision. This adjustment significantly reduced errors and improved the overall quality of annotations.

The resulting Bangla OCR corpus presented in this paper is the largest and most diverse dataset available for the Bangla language. Its comprehensive coverage of various document types and sources makes it a valuable resource for OCR research and development. Researchers can categorize the corpus based on document type, words, and characters to suit their specific needs. This corpus will be an invaluable asset to the Bangla language research community. We anticipate that the research community will leverage this corpus to advance the field of Bangla OCR, leading to enhanced language understanding and accessibility in the digital era. Its availability will enable advancements in Bengali OCR technology, serving as a benchmark for evaluating and improving OCR algorithms. Moreover, it will facilitate the training and testing of OCR models designed explicitly for the Bangla language.

Ethics Statement

All the human resources used in this data collection process are well-remunerated. Data processing Engineers (DPE) and Research Assistants (RA) were Contractual Basis employees during the data collection process. When writing the Handwritten data, text scripts are collected from open-source books with the proper consent of the original author and publisher. Also, the writers' information like Name, Age, Eduacation, and Place of residence are collected for internal data collection tracking, and this information will be anonymized and will be shared for forensic use. All personal information is excluded from the final dataset for every data type, including Handwritten or Printed documents. All printed data sources, including Computer Compose, Letterpress, and Typewriters documents used for creating this dataset, are the properties of The ICT Division, Government of the People's Republic of Bangladesh, and will be published by the ICT Division for facilitating further research on Bangla Language.

Acknowledgements

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. We acknowledge the support from *Apurba-DIU Research Lab (ADRL), DIU NLP and ML Research Lab*, and *Department of Computer Science and Engineering, Daffodil International University* for the continuous support throughout the journey.

References

Saad Bin Ahmed, Saeeda Naz, Salahuddin Swati, and Muhammad Imran Razzak. 2019. Handwritten urdu character recognition using one-dimensional blstm classifier. *Neural Computing and Applications*, 31:1143–1151.

AKM Shahariar Azad Rabby, Sadeka Haque, Shammi Akther Shahinoor, Sheikh Abujar, and Syed Akhter Hossain. 2018. A universal way to collect and process handwritten data for any language. *Procedia Computer Science*, 143:502–509. 8th International Conference on Advances in Computing & Communications (ICACC-2018).

Mithun Biswas, Rafiqul Islam, Gautam Kumar Shom, Md. Shopon, Nabeel Mohammed, Sifat Momen, and Anowarul Abedin. 2017. Banglalekha-isolated: A multi-purpose comprehensive dataset of handwritten bangla isolated characters. *Data in Brief*, 12:103–107.

Avishek Das, AKM Shahariar Azad Rabby, Ibna Kowsar, and Fuad Rahman. 2022. A deep learning-based unified solution for character recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1671–1677. IEEE Computer Society.

Iyad Abu Doush, Faisal AIKhateeb, and Anwaar Hamdi Gharibeh. 2018. Yarmouk arabic ocr dataset. *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, pages 150–154.

Jannatul Ferdous, Suvrajit Karmaker, AKM Shahariar Azad Rabby, and Syed Akhter Hossain. 2021. Matrivasha: A multipurpose comprehensive database for bangla handwritten compound characters. In *Emerging Technologies in Data Mining and Information Security*, pages 813–821, Singapore. Springer Singapore.

Donatella Firmani, Paolo Merialdo, Elena Nieddu, and Simone Scardapane. 2017. In codice ratio: Ocr of handwritten latin documents using deep convolutional networks. In *AI*CH@AI*IA*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Alex Graves and Jürgen Schmidhuber. 2008. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Sadeka Haque, AKM Shahariar Azad Rabby, Md Sanzidul Islam, and Syed Akhter Hossain. 2018. Shonkhanet: A dynamic routing for bangla handwritten digit recognition using capsule network. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 159–170. Springer, Singapore.

Md Majedul Islam, Avishek Das, Ibna Kowsar, AKM Shahariar Azad Rabby, Nazmul Hasan, and Fuad Rahman. 2021. Towards building a bangla text recognition solution with a multi-headed cnn architecture. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1061–1067. IEEE.

Martin Kišš, Michal Hradiš, and Oldřich Kodym. 2019. Brno mobile ocr dataset. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1352–1357. IEEE.

Vaibhav. V. Mainkar, Jyoti A. Katkar, Ajinkya B. Upade, and Poonam R. Pednekar. 2020. Handwritten character recognition to obtain editable text. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 599–602.

Urs-Viktor Marti and H. Bunke. 2002. The iam-database: An english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

AKM Shahariar Azad Rabby, Sadeka Haque, Sheikh Abujar, and Syed Akhter Hossain. 2018. Ekush-net: Using convolutional neural network for bangla handwritten recognition. *Procedia computer science*, 143:603–610.

AKM Shahariar Azad Rabby, Sadeka Haque, Md. Sanzidul Islam, Sheikh Abujar, and Syed Akhter Hossain. 2019. Ekush: A multipurpose and multitype comprehensive database for online off-line bangla handwritten characters. In *Recent Trends in Image Processing and Pattern Recognition*, pages 149–158, Singapore. Springer Singapore.

AKM Shahariar Azad Rabby, Md. Majedul Islam, Nazmul Hasan, Jebun Nahar, and Fuad Rahman. 2021. Borno: Bangla handwritten character recognition using a multiclass convolutional neural network. In *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 1*, pages 457–472, Cham. Springer International Publishing.

Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. 2020. Roadtext-1k: Text detection & recognition dataset for driving videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11074–11080. IEEE.

Md Jamiur Rahman Rifat, Mridul Banik, Nazmul Hasan, Jebun Nahar, and Fuad Rahman. 2021. A novel machine annotated balanced bangla ocr corpus. In *Computer Vision and Image Processing*, pages 149–160, Singapore. Springer Singapore.

G. Abdul Robby, Antonia Tandra, Imelda Susanto, Jeklin Harefa, and Andry Chowanda. 2019. Implementation of optical character recognition using tesseract with the javanese script target in android application. *Procedia Computer Science*, 157:499–505. The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019) : Enabling Collaboration to Escalate Impact of Research Results for Society.

Koushik Roy, Md. Sazzad Hossain, Pritom Saha, Shadman Rohan, Imranul Ashrafi, Ifty Mohammad Rezwan, Fuad Rahman, B. Hossain, Ahmedul Kabir, and Nabeel Mohammed. 2023. A multifaceted evaluation of representation of graphemes for practically effective bangla ocr. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–23.

Md Istiak Hossain Shihab, Md Rakibul Hasan, Mahfuzur Rahman Emon, Syed Mobassir Hossen, Md Nazmuddoha Ansary, Intesur Ahmed, Fazle Rabbi Rakib, Shahriar Elahi Dhruvo, Souhardya Saha Dip, Akib Hasan Pavel, et al. 2023. Badlad: A large multi-domain bengali document layout analysis dataset. In *International Conference on Document Analysis and Recognition*, pages 326–341. Springer.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*.

Ying Zhang, Lionel Gueguen, Ilya Zharkov, Peter Zhang, Keith Seifert, and Ben Kadlec. 2017. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In *SUNw: Scene Understanding Workshop - CVPR 2017*, Hawaii, U.S.A.

A Appendix

The datasets have images from different categories, including computer-composed words (isolated + running), computer-composed characters (isolated + running), typewriter-composed paper words (isolated + running), typewriter-composed paper characters (isolated + running), letterpress-composed words (isolated + running), letterpress-composed characters (isolated + running), offline handwriting isolated characters, offline handwriting isolated words, offline handwriting running words, outdoor words, and dynamic handwritten data. Each category is balanced based on factors such as font type, size, noise, source, age, education, place, light, and color. The table in the appendix displays category-wise data statistics. Table 2 shows data distribution statistics before and after annotation.

The dataset has a wide range of words, grapheme, and character distribution. Table 3 shows the distribution of unique words, grapheme, and characters presented in the dataset.

Table 2: Category wise data statistics

Category Name	**Data before annotation**	**Data after annotation**	**Balanced by**
Computer Compose words (Isolated + Running)	6,03,520	5,42,596	Font type, Size, Noise and Source
Computer Compose character (Isolated + Running)	2,33,859	1,78,158	Font type, Size, Noise and Source
Typewriter Composed, Paper words (Isolated + Running)	6,76,841	2,00,646	Font type, Size, Noise and Source
Typewriter Composed, Paper characters (Isolated + Running)	96,202	50,686	Font type, Size, Noise and Source
Letterpress Composed words (Isolated + Running)	9,56,518	4,12,248	Font type, Size, Noise and Source
Letterpress Composed character (Isolated + Running)	1,47,786	84,295	Font type, Size, Noise and Source
Offline Handwriting Isolated Character	59,805	57,319	Age and Education
Offline Handwriting Isolated Word	5,88,158	5,10,182	Age and Education
Offline Handwriting Running Word	11,06,769	10,27,423	Age and Education
Outdoor word	79,207	52,520	Place, Light and Color
Dynamic Handwritten		10,00,000	Font type and Size

Table 3: Category-wise distribution of unique Words, Graphemes, and Characters

Category Name	**Unique Character**	**Unique Grapheme**	**Unique Words**
Computer Compose	192	541	58432
Typewrite	178	406	22181
Letterpress	214	536	62144
Offline Handwriting	216	614	55063
Outdoor			11349
Dynamic Handwriting			45324