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

Hasmot Ali, AKM Shahariar Azad Rabby, Md. Majedul Islam, A.K.M Mahamud, Nazmul Hasan, Fuad Rahman
Apurba Technologies, Sunnyvale, California, USA

Abstract

This study introduces the most extensive gold standard corpus for Bangla characters and words, comprising over 4 million human-annotated images and 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.

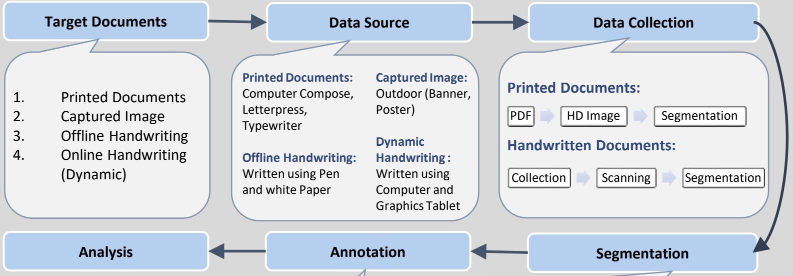
Background

The research and development for low resource language like Bangla is a quite far away from the advancement of current technology. Technology like Optical Character Recognition is not very useful because of its terrible performance for Bangla Language. The reason behind this is the complexity of **Bangla Text Structure**, different **Documents Type** and **unavailability of a dataset**. There are some available dataset only for printed character, words, digit and some domain specific words which can improve the OCR performance for a very limited field. There is also a synthetic dataset of 1 million words is available for Bangladesh. A lot of offline handwriting dataset is presented but there is no dynamic handwriting data is available so far.

Objectives

- ❖ To fasten the research progress of Low Resource Language
- ❖ To improve the accuracy of Bangla Language
- ❖ To collect a gold standard dataset for diverse documents types
- ❖ To provide a word level and character level dataset

Materials and Methods



Target Documents

1. Printed Documents
2. Captured Image
3. Offline Handwriting
4. Online Handwriting (Dynamic)

Data Source

Printed Documents: Computer Compose, Letterpress, Typewriter

Captured Image: Outdoor (Banner, Poster)

Offline Handwriting: Written using Pen and white Paper

Dynamic Handwriting : Written using Computer and Graphics Tablet

Data Collection

Printed Documents: PDF → HD Image → Segmentation

Handwritten Documents: Collection → Scanning → Segmentation

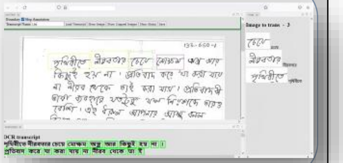
Segmentation

Annotation

Analysis

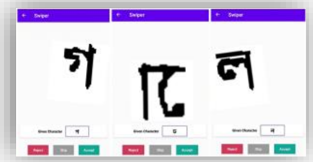
Segmentation:

- ❖ **Pixel-wise Semantic Segmentation** approach is applied for Computer Compose, Letterpress and Typewriter documents. The approach was started with identifying lines within documents, segmenting the line from documents and then segmenting words from the line.
- ❖ Web based **Cropping Tools** was used for Handwriting and Outdoor images. Designed for cropping words manually from documents and tagging with associated words.



Annotation:

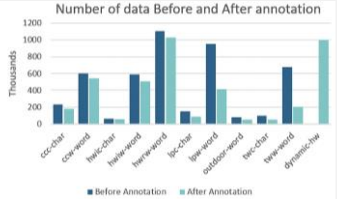
- ❖ An **Android Application** was initially used to annotate words and character having a swiper like interface.
- ❖ **Annotation Web Tools** is designed for tagging the corresponding words or character with the corresponding images. Every single data is annotated by 3 individuals annotators and validated by one supervisor level annotators, responsible for validating previous 3 annotators agreements to ensure gold standard criteria.



Results

Statistics:

The Dataset is well balanced with words, character and grapheme. The dataset have **4 million words and character** with **1 million dynamic handwriting** data with coordinate. The amount of Typewriter and Letterpress data is less due to the unavailability of these documents.



Kappa Score:

Kappa coefficient assesses the inter-annotator agreement among two or more annotators. A higher Kappa Score indicates a higher level of agreement between annotators. In our case, we obtained a Kappa Score of **0.91 for computer-composed**, **0.93 for Letterpress**, and **0.78 for Typewriter** documents.

CRNN-VDS Model Performance:

A small subset of this data is provided to a R&D lab for evaluating the dataset. They have tested their CRNNVDS model with VDS Character Representation. CRNN-VDS is trained on a large-scale synthetic dataset having 2 million samples.

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

Conclusion

- ❖ A collection of six million human annotated data
- ❖ Six different types of documents
- ❖ Balanced by Character, Words and Grapheme
- ❖ Every data point is checked by 4 individuals
- ❖ One million of dynamic handwritings data with coordinate value