Project ID: 35

Optical Character Recognition for bangla handwritten and printed documents

(Single Grapheme Prediction based model, Bangla HOCR)

অভিযাত্রিক

Md. Nazmuddoha Ansary

Syed Mobassir Hossain

Md. Aminul Islam

In association with:-

APSIS Solutions LTD

Bengali.ai

কাজের বর্ণনাঃ

- বাংলা গ্রাফিম ব্যবহারের মাধ্যমে ছাপা ও হাতে লেখা শব্দের জন্য ও.সি.আর. তইরি করা।
- বর্তমানে সহজলভ্য ও বহুলব্যবহৃত ও.সি.আর. ব্যবস্থা (যেমনঃ টেসারেক্ট বা ইজি ও.সি.আর.) গুলি ছাপা লেখার জন্য কাজ করলেও হাতে লেখা ছবির জন্য তেমন ভাল কাজ করে না।
- আমাদের কাজের মূল অংশ হল হাতে লেখা শব্দ গুলোকে একটি নির্দিষ্ট রূপে রূপান্তরের মাধ্যমে সহজলভ্য ও বহুলব্যবহৃত ও. সি.আর. ব্যবস্থা গুলিকে হাতে লেখা শব্দ পড়ানোর সক্ষমতা দেয়া।
- এতে করে হাতে লেখা ও ছাপা লেখার জন্য আলাদা আলাদা ব্যবস্থার দরকার হবে না।

Live Demo

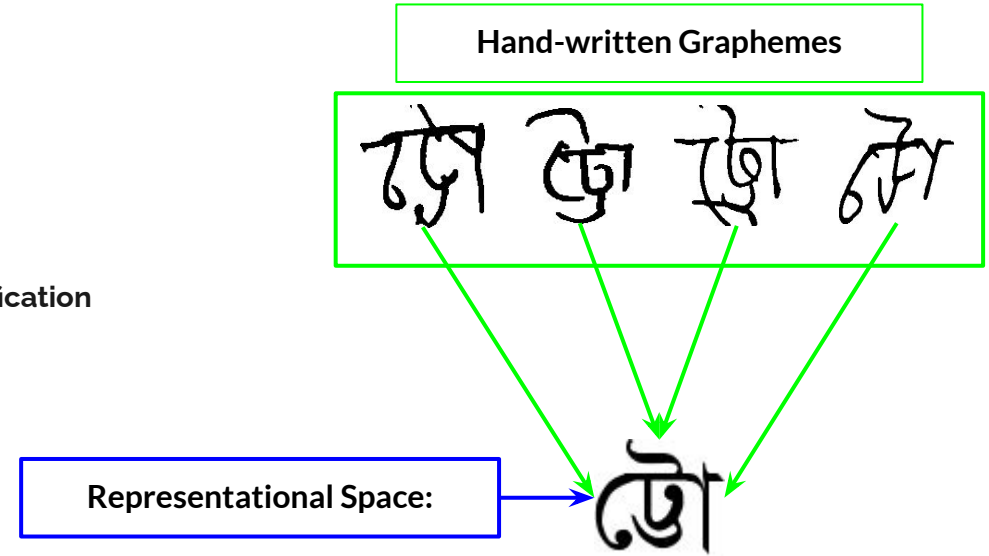
Resources

- Data preparation: https://github.com/mnansary/banglaOCR
- Model inference:https://github.com/mnansary/bnhocr

**Grapheme: the smallest meaningful contrastive unit in a writing system.

Focus

- Bangla HOCR
  - Printed Documents (ALMOST Solved)
  - Handwritten Character and/Or Grapheme Classification
- Word/Line Based Approach:
  - Sequential Detection Based
  - Receptive Field Based
  - 13-14 SOTA



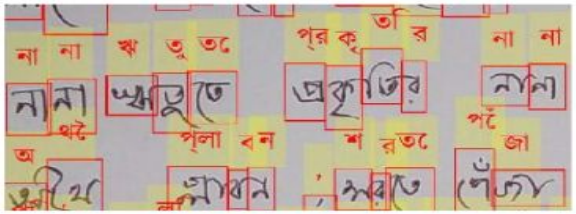
The Idea:

- Modifying each grapheme to one specific representation
  - NOT Classification
  - NOT GAN/Generative Models
  - Character/Unicodes are not suitable

Significance:

- Is not bound to grapheme-level only
- Representational spaces:
  - Font Faced (Easiest)
  - Gaussian Heatmap
  - Fourier Contour
  - Positional Encoding
  - Skeletal Encoding
  - And many more..
- Adaptable with known methods

1. HOCR with Sequential Detection



Images from Nishatul et.al(2019) [link:10.1109/ICDAR.2019.00045 (Segmentation-Free Bangla Offline Handwriting Recognition using Sequential Detection of Characters and Diacritics with a Faster R-CNN)

- Character and Diacritics NOT Grapheme

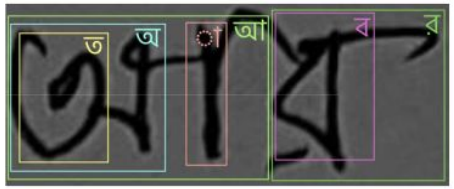
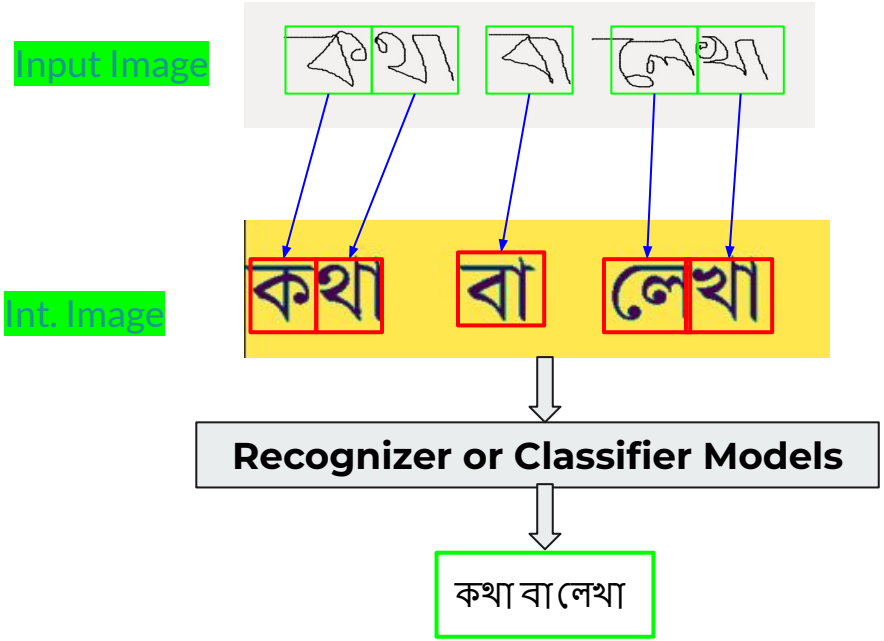


Fig. 6. Sample of the detection overlap issue. Green boxes are the proper detection and the others are detected look-alike sub-characters.

Adaptable Idea

- Grapheme based Detection (solves overlapping)
- (CLASSIFICATION):Huge Number of grapheme classes (popular:1300, possible: ~17000)
- Opens the possibility to use existing Recognizer like tesseract
- **Demo:** https://www.youtube.com/watch?v=7Ye9oV1MqbU



2. HOCR with CRNN-CTC

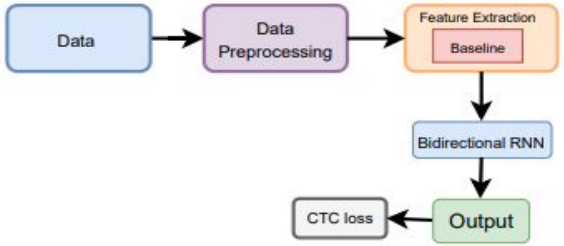
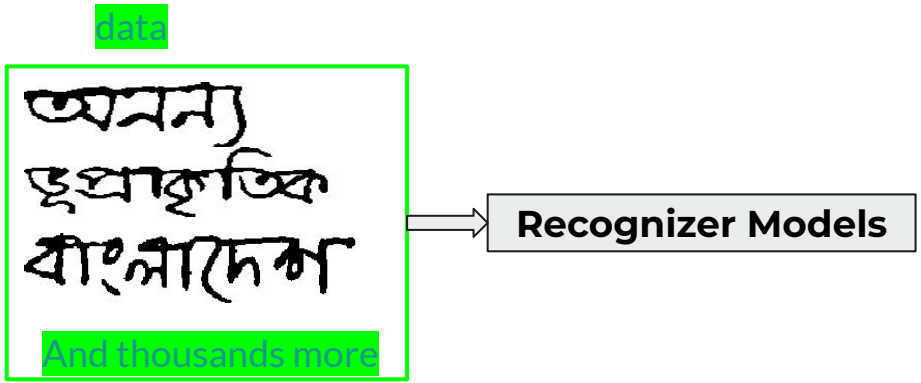
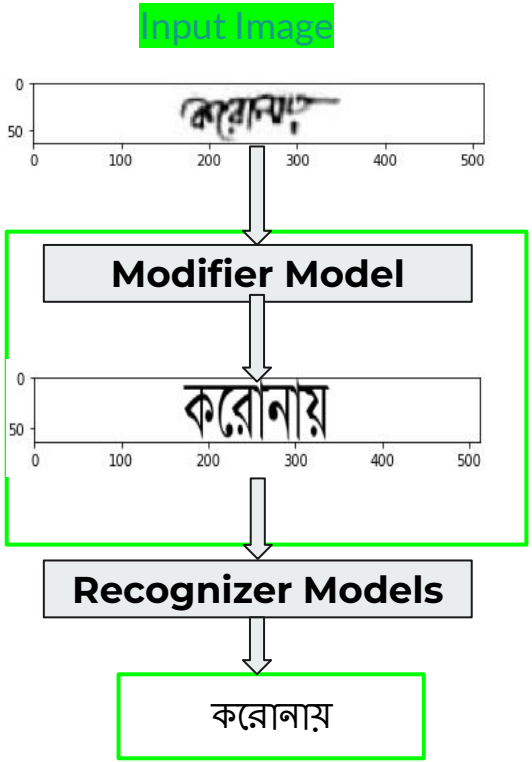


Image from Farisa et.al(2021) [link:https://arxiv.org/pdf/2105.04020.pdf](End-to-End Optical Character Recognition for Bengali Handwritten Words)



Adaptable Idea(NBA)

- Convert written text to a “Typed Font Faced” image
- Re-use available recognition systems



Thank You

Questions?