6-19-2018

Handwriting Recognition of Bangla and Similar Scripts

Nishatul Majid
Boise State University

Handwriting Recognition of Bangla and Similar Scripts

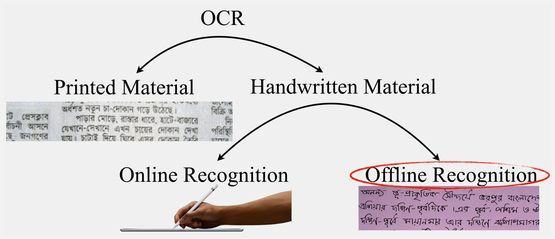
BOISE STATE UNIVERSITY

by Nishatul Majid (নিশাতুল মজিদ), Advisor - Dr. Elisa Barney Smith

Department of Electrical and Computer Engineering

I. Objective

- The idea is to recognize text from document Images. This is known as Optical Character Recognition (**OCR**).



- The target script is **Bangla** and **other Indic scripts** which share many similarities.

II. About Bangla

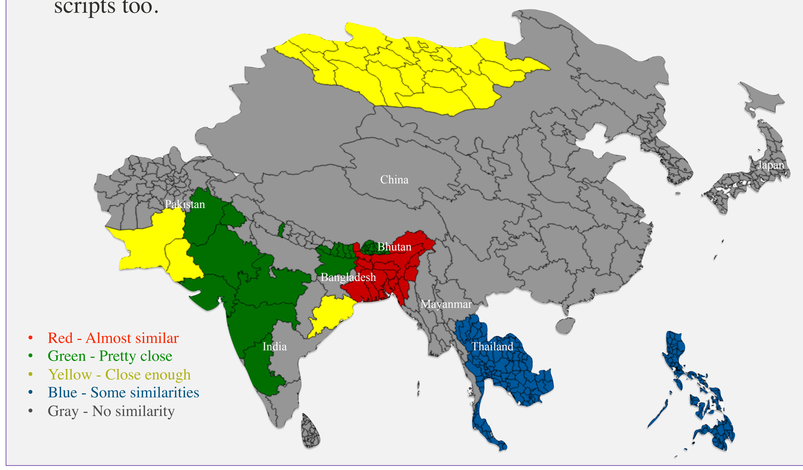
- Bangla a.k.a Bengali is the **7th most spoken native language** in the world, spoken by over 205 million of people.
- The Bangla script, along with the Assamese, is the **5th most widely used writing system** in the world.
- It is the national and official language of the People's Republic of **Bangladesh**, and official language of several states in India such as **West Bengal**, **Tripura**, **Assam**, **Andaman** etc.
- Bangla is written from left to right with no upper or lower cases.
- The script consists of 11 vowels, 10 vowel diacritics, 39 consonants, several hundred consonant conjuncts, more than 10 consonant diacritics, 10 numbers and several punctuation marks.
- Words are connected by a distinctive horizontal line along the top of the characters in a word called the "**Matra**".

III. Potential for this Research

- **Task Automation** - such as interpretation of handwritten addresses, bank checks, flyers, posters, notices, banners, invitation cards, tickets, answer sheets from exams etc.
- Fast and effective **archival of handwritten documents**, such as literature, journals, academic notes, etc.
- Search handwritten documents with codewords.
- Signature, writer verification.
- Assistive technology for blind or visually impaired people.

IV. Similar Scripts

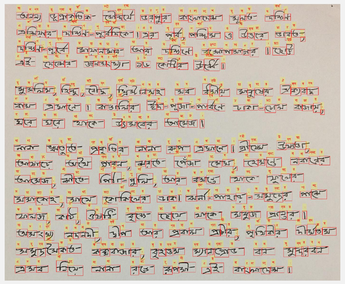
- Bangla is an **Indo Aryan Language**. Its script originated from the **Brahmic** family, written in the **Abugida** writing system.
- Scripts originating from the same class share striking similarities with Bangla. Scripts such as **Devanagari**, **Assamese**, **Gurmukhi**, **Gujarati,** etc.
- Devanagari is used for over 120 languages including Hindi, Marathi, Nepali etc. It is the **4th most widely used writing system** in the world.
- Recognition methods for Bangla are expected to work pretty well for these scripts too.

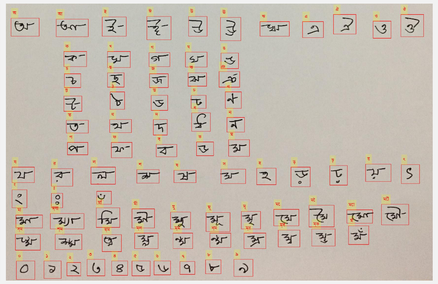


V. The BSU Bangla Dataset

Advancement in research and development related to machine learning is highly dependent on the availability of **good** datasets. Here, we present the BSU Bangla Dataset aiming to foster recognition progress.

- Handwriting samples from more than **100 participants**. Each provided us **an essay** and **a page of isolated characters**.
- This essay was carefully designed so that it contains almost all basic characters, all possible vowel diacritics and several high frequency conjuncts.
- All the pages were **tagged with associated ground truth** for characters, words and lines.

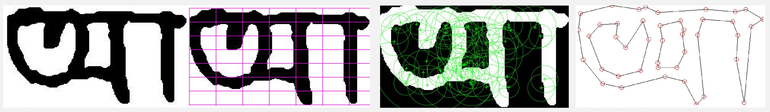




VI. Basic Character Recognition

We developed a basic character (isolated) recognition module. This involves -

- **Preprocessing -**
  - Binarization using Otsu's threshold method
  - Fixing sizes at 128 pixels of height and
  - Applying 2D Gaussian smoothing filter
- **Feature extraction -**



- **Classification** - with Support Vector Machine based on a **cubic kernel.**
- **Results -**

Training Dataset	Testing Dataset	Previously Reported Max Accuracy	Our Recognition Results
CMATERdb 3.1.2	CMATERdb 3.1.2	86.40% - by Roy et al. [1]	92.87%
	BSU db	n. a.	91.39%
ISI db	ISI db	95.84% - by Bhattacharya et al. [2]	93.10%
	BSU db	n. a.	89.24%
BanglaLekha db	BanglaLekha db	95.10% - by Alif et al. [3]	96.80%
	BSU db	n. a.	95.78%
Combined	BSU db	n. a.	96.42%

VII. Conclusion

The plans for future work includes -

- Develop an **extensive character recognition** model which can identify the consonant conjuncts and the vowel diacritics along with the basic characters. A **multistage hierarchical classification** architecture can be a promising candidate for this.
- **Segment words into characters**. Several researchers are exploring this. **Water reservoir** based algorithms have good potential to solve this.
- **Segmentation free recognition** approaches are also very promising and trendy these days; a word is recognized as a whole rather than identifying its component characters. Statistical modeling such as **Hidden Markov Models** are very powerful tools for these approaches.

References:

[1] A. Roy, N. Das, R. Sarkar, S. Basu, M. Kundu, and M. Nasipuri, "Region selection in handwritten character recognition using artificial bee colony optimization," in *Emerging Applications of Information Technology (EAIT)*, 2012 Third International Conference on. IEEE, 2012, pp. 183–186.

[2] U. Bhattacharya, M. Shridhar, S. K. Parui, P. K. Sen, and B. B. Chaudhuri, "Offline recognition of handwritten Bangla characters: an efficient two-stage approach," *Pattern Analysis and Applications*, vol. 15, no. 4, pp. 445–458, jun 2012.

[3] M. A. R. Alif, S.Ahmed, and M.A.Hasan,"Isolated Bangla handwritten character recognition with convolutional neural network," in *Computer and Information Technology (ICCIT)*, 2017 20th International Conference of. IEEE, 2017, pp. 1–6.