8-7-2018

Introducing the Boise State Bangla Handwriting Dataset and an Efficient Offline Recognizer of Isolated Bangla Characters

Nishatul Majid
Boise State University

Elisa H. Barney Smith
Boise State University

Introducing the Boise State Bangla Handwriting Dataset and an Efficient Offline Recognizer of Isolated Bangla Characters

BOISE STATE UNIVERSITY

by Nishatul Majid (নিশাতুল মজিদ) and Dr. Elisa H. Barney Smith (ডঃ এলিসা এইচ. বার্নি স্মিথ)

Department of Electrical and Computer Engineering, Boise State University, Boise, Idaho, USA.

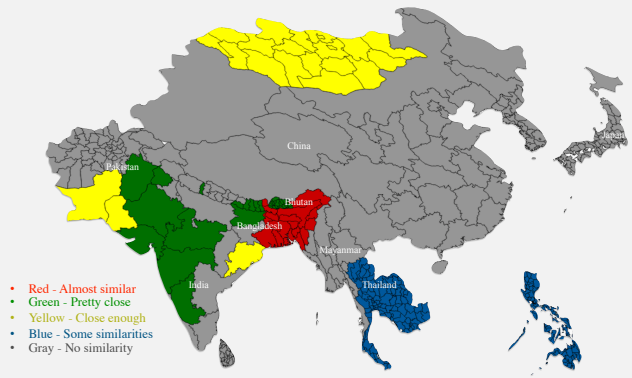
I. Abstract

- This introduces the **Boise State Bangla Handwriting Dataset**, a publicly accessible offline handwriting dataset of Bangla script. This can be found at *https://scholarworks.boisestate.edu/saipl/1/*
- **A basic character recognition method** is presented where the features are extracted based on zonal pixel counts, structural strokes and grid points with U-SURF descriptors modeled with bag of features.
- Benchmarking with this approach on 3 other publicly available Bangla datasets is reported. The highest classification accuracy obtained with a**n SVM classifier based on a cubic kernel is 96.8%.**

Scan me

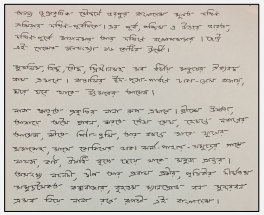
II. About Bangla and Similar Scripts

- Bangla a.k.a Bengali is an Indo-Aryan language. Its script originated from the **Brahmic** family, written in the **Abugida** writing system.
- It 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.
- Many scripts such as **Devanagari**, **Assamese**, **Gurmukhi**, **Gujarati,** etc. share striking similarities with Bangla. Scripts. Hence, the character recognition approach is expected to work with these scripts as well.



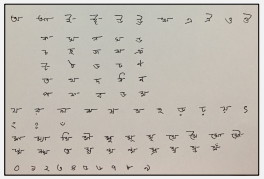
III. The BSU Bangla Dataset

Camera based acquisition of handwriting samples from **100 participants**.



Essay Pages -

- 104 words or 364 characters
- 49 basic characters, all 11 vowel diacritics and 32 high frequency conjuncts.



Isolated Character Pages -

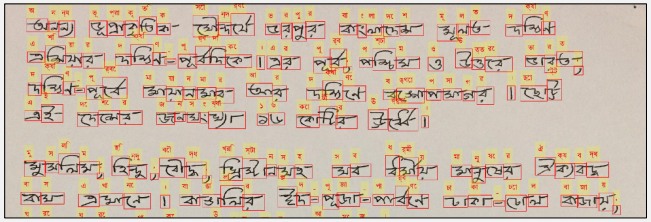
- 84 isolated units
- All basic characters, vowel diacritics and several high frequency conjuncts.

Tagging:

- At line, word and character level with ground truth

```
Line001 Word001 Char001 অ – 115,70,94,57
Line001 Word001 Char002 ন – 213,80,45,47
Line001 Word001 Char003 ন্য – 257,79,69,73
Line001 Word002 Char004 ভূ – 390,79,78,72
Line001 Word002 Char005 প্রা – 466,80,75,63
Line001 Word002 Char006 কৃ – 547,79,56,72
Line001 Word002 Char007 তি – 608,61,75,76
```

```
Char012 ক – 703,313,140,95
Char013 খ – 970,313,140,101
Char014 গ – 1224,320,127,88
Char015 ঘ – 1446,317,133,94
Char016 ঙ – 1664,323,107,98
Char017 চ – 706,454,101,87
Char018 ছ – 970,447,101,91
```



Demographic Information:

- Age
- Gender
- Profession
- Left/Right handedness

IV. Isolated Character Recognition

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

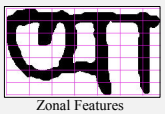
Preprocessing -

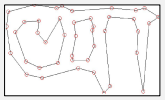
- Binarization using Otsu's threshold method
- Normalizing height to 128 pixels
- Applying 2D Gaussian smoothing filter

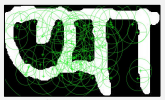

Preprocessed Image

Feature extraction -

- **Zonal Features:** The image was split in 8X8 zones. The ratios of white and black pixel in each zone were formed into a 64-bit feature-vector.
- **Pattern Features:** The stroke direction and length were extracted from the exterior of the character images. Later these were quantized to obtain a 67-bit feature vector.
- **Gradient Features:** Upright Speed Up Robust Features (U-SURF) from a 8X8 grid intersection was submitted to a Bag-of-features model to obtain a 500 word visual vocabulary.


Zonal Features


Stroke Pattern Features


U-SURF Features

Classification -

Support Vector Machine based on a **cubic kernel** with a **One Versus One (OVO)** multi-class scheme.

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

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, 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.

V. Future Plans

- Continue **growing the Boise State Handwriting Dataset** with more samples.
- Develop an **extensive character recognition** module to include the consonant conjuncts and the vowel diacritics along with the basic characters.
- Develop a **Segmentation free recognition** module and compare with the existing segmentation based approaches using the dataset.
- Acquire a flatbed scanner image of samples to compliment the camera-based.