Handwriting Recognition of Bangla and Similar Scripts

Nishatul Majid
Boise State University
Handwriting Recognition of Bangla and Similar Scripts

Abstract
This research is about offline Bangla handwriting text recognition. Here we introduce a publicly accessible dataset, as well as a basic character recognition scheme. The dataset contains pages with a 104 word essay and a collection of 84 isolated alpha-numeric characters. All the components in the pages are tagged with the associated ground truth information. The character recognition scheme presented here uses zonal pixel counts, structural strokes and bag of features modeled with grid points using U-SURF descriptor as features. The maximum classification accuracy we obtain is 96.8% using an SVM classifier with a cubic kernel.

Disciplines
Computer Sciences | Electrical and Computer Engineering

This scholarly poster is available at ScholarWorks: https://scholarworks.boisestate.edu/gss_2018/54
Handwriting Recognition of Bangla and Similar Scripts
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 -

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Testing Dataset</th>
<th>Previously Reported Max Accuracy</th>
<th>Our Recognition Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMAterdam 3.1.2</td>
<td>BSU db</td>
<td>96.42%</td>
<td>92.87%</td>
</tr>
<tr>
<td>CMAterdam 3.1.2</td>
<td>ISI db</td>
<td>95.78%</td>
<td>91.39%</td>
</tr>
<tr>
<td>BSU db</td>
<td>BSU db</td>
<td>95.84%</td>
<td>92.87%</td>
</tr>
<tr>
<td>SQL db</td>
<td>BSU db</td>
<td>93.10%</td>
<td>91.39%</td>
</tr>
<tr>
<td>Combined</td>
<td>BSU db</td>
<td>n. a.</td>
<td>96.42%</td>
</tr>
</tbody>
</table>

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: