Numta - Assembled Bangla Handwritten Digits

Samiul Alam[1], Tahsin Reasat [1], Rashed Doha [2], Ahmed Imtiaz Humayun [1],
[1]*Department of EEE*, *Bangladesh University of Engineering & Technology*, Dhaka, Bangladesh
[2]*Department of Mechanical Eng*, *Bangladesh University of Engineering & Technology*, Dhaka, Bangladesh

***Abstract*—Hand written digits are the basic starting point for a complete optical character recognition system. If a system can detect hand written numbers, then immediately it has a range of possibilities for practical applications. It also provides important insight on how to extend the system to recognize characters and scopes for improvement. So, with a vision to create a complete functioning OCR for Bengali language, we have assembled a large dataset (85,000+ images) of hand-written Bengali digits collected from various sources. The goal of this data-set is to create a Bengali digit recognizer capable of high accuracy and noise tolerance. The data-set is meant to be the start of a series of datasets aimed at furthering natural language processing research on Bengali.**

***Index Terms*—Bangla Digits, OCR, Bangla MNIST**

I. Introduction

This paper documents the process and salient statistics of the assembled Bengali hand written digit dataset. We have assembled data from several sources and combined them all under strict guidelines of legibility so that there is minimum noise in the training data. The train-test split was also done with strict consideration given to minimizing correlation between training and testing data.

II. Dataset Description

The dataset is a combination of six datasets that were gathered from different sources and at different times. However each of them were checked rigorously under the same evaluation criterion so that all digits were at least legible to a human being without any prior knowledge. Descriptions of these datasets including collection methodology, image segmentation and extraction and image formats of these datasets are described in this section. The information regarding contributors and image collection is summarized in Table II, Table I, and Table III.

A. Bengali Handwritten Digits Database (BHDDB)

BHDDB dataset was collected from students in the Department of Computer Science and Engineering of Bangladesh University of Engineering and Technology. The students were given a form to write down the numerals. The forms had regular grid pattern in which the numerals were inscribed. The forms were scanned in color. The forms had a marker on each of its four corners which were used to align the image borders with the grid lines. The digit extraction procedure is described in Section III.

B. BUET101 Database (B101DB)

The participants of this dataset wrote the digits on a paper which was scanned at 600 dpi. The digits were than manually cropped and labeled.

C. OngkoDB

OngkoDB was collected from a group of students from the Department of Computer Science and Engineering of Bangladesh University of Engineering and Technology. They filled up forms which did not have markers on the corners. The forms were scanned in gray-scale. Since the forms had no markers, a different extraction approach was taken. The images of the digits were extracted by first re-orienting feature points of SURF(Speeded Up Robust Feature) of the original image to a reference image and then extracting all images of the digits. The automated extraction procedure was not fully accurate and the dataset went through rigorous pruning (Details in Section III).

D. ISRTHDB

ISRTHDB was collected from students in Institute of Statistical Research and Training, Dhaka University. The collection process and evaluation followed here was done in the same format as BHDDB dataset. The data for this was collected after BHDDB and with strong collaboration with people involved in the former. As such, the raw data of this dataset is much cleaner than its predecessor.

E. BanglaLekha-Isolated Numerals

BanglaLekha-Isolated [**?**] dataset contains Bangla handwritten numerals, basic characters and compound characters. The data was collected from literate native bengali speakers from Dhaka and Comilla. The digits in this dataset contained erroneous labels and outliers which were cleaned and included in our dataset. The Banglalekha-Isolated dataset were released as preprocessed binary images. According to the authors, the following preprocessing steps were taken:

- Foreground and background were inverted so that images have a black background with the letter drawn in white.
- Noise removal was attempted by using the median filter.
- An edge thickening filter was applied.
- Images were resized to be square in shape with appropriate padding applied to preserve the aspect ratio of the drawn character.

F. UIUDB

UIUDB dataset was collected by students of United International University from scanned documents, windows paint images and cell phone camera photos. Due to the nature in which data was gathered, this dataset is the hardest to train on and we have left it only in test set.

III. Extraction Procedures

The large majority of the data excluding Banglalekha Dataset was extracted following one of two algorithms depending on whether markers were present in the forms. The resulting extracted data and the corresponding problems in each algorithm have been illustrated in the following sub sections.

A. Marker based alignment and Grid Detection

In case of BHDDB and ISRTDB, the raw scans had square markers. The forms had four markers placed at four edges of the region of interest which contain a rectangular grid like table. Digits were hand written inside the grids. The raw images are denoted by I^R such that $I^R \in \mathbb{R}^{W,H}$. Here, W and H is the dimension of the image. The images were transformed into binary images I^B and passed through a blob detection algorithm and segmented into blobs. The set of blobs is denoted as $\mathcal{B}$. For each blob $b \in \mathcal{B}$, perimeter P_b, Area A_b, centroid CE_b, and bounding box (W_b, H_b) were measured. Two properties circularity C_b and extent E_b were then defined as follows:

$$C_b = \frac{P_b^2}{4\pi A_b} \tag{1}$$

$$E_b = \frac{A_b}{W_b H_b} \tag{2}$$

The possible centroids of the markers, were determined by the segmented areas that satisfied the conditions defined as:

$$1.1 \leq C_b \leq 1.6 \tag{3}$$

$$E_b \geq 0.5 \tag{4}$$

The set of centroids that fulfill the condition is denoted by $\mathcal{V}$. If three marker centroids could be determined, the image was transformed and cropped to a rectangle whose vertices lies on those centroids. If more than three were found, then the four centroids that formed the rectangle with height and width closest to a reference rectangle were picked and the image was transformed accordingly. The reference rectangle was defined using the dimensions (height H and width W) of the raw image. Then the horizontal and vertical lines of the grid would be in alignment with the axes. The aligned images (I^A) were then summed in each direction separately which outputs two arrays.These arrays would have strong peaks along grid lines. So by using peak detection, all intersection points of grid lines were determined.

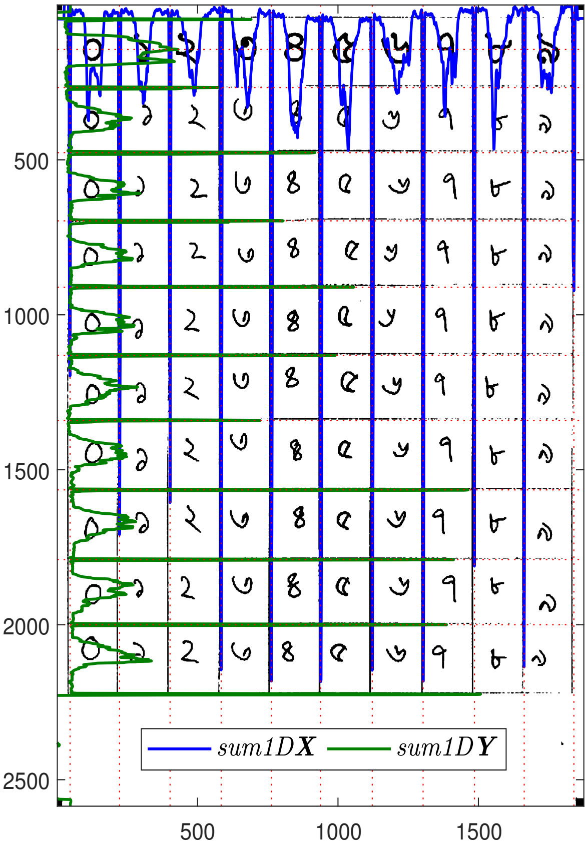


Fig. 1. Summation of the aligned image along each axis creates two one dimensional signals with distinct peaks (shown in green and blue).

By using the points, the crops of the hand written digits (d) were extracted and included into the set of extracted digits denoted by $\mathcal{D}$. The digit extraction algorithm is illustrated in Algorithm 1. Since there was no margin between each grid box, some of the images had extensions from adjacent boxes intruding in their box (Fig. 2(e)). These were manually sorted out later.

B. Markerless Alignment with SURF and Square Detection

In case of OngkoDB, there was no marker. So an empty form was used as reference image which was perfectly aligned and using Speeded Up Robust Features, we realigned all the scanned image to that reference image.

We created a second reference image where all the boxes were filled in. Then the centroids of the squares in this image were used to cut out boxes of width and height slightly larger than the box containing the digits so that the cropped image has both the hand written digit and the bounding box.
Then the pixels were summed up in each direction separately which outputs two arrays. These arrays would have strong peaks along bounding box lines if they exist. So by using peak

Algorithm 1: Digit extraction from marker based forms

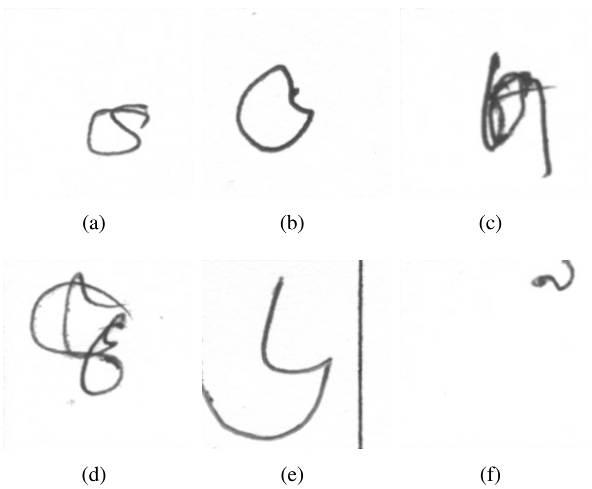
```
Input  : Raw Scan of image, $\{I^R_{i,j}\}^{W,H}_{i,j=1}$
Output: Cropped set of digits, $\mathcal{D}$
Initialize: $\mathcal{V} \leftarrow \{\phi\}, \mathcal{D} \leftarrow \{\phi\}$
Pre-Processing:
$I^B \leftarrow$ Binarize($I^R$)
$I^B \leftarrow$ MedianFilter($I^B$, window = 5 × 5)
$\mathcal{B} \leftarrow$ BlobDetector($I^B$)
if $|\mathcal{B}| \leq 2$ then
    $I^B \leftarrow$ MedianFilter($I^B$, window = 15 × 15)
    $\mathcal{B} \leftarrow$ BlobDetector($I^B$)
end
Determine rectangle vertices:
for $\forall b \in \mathcal{B}$ do
    $CE_b, P_b, A_b, W_b, H_b \leftarrow$ BlobProperties($b$)
    $C_b \leftarrow \frac{P_b^2}{4\pi A_b}$
    $E_b \leftarrow \frac{A_b}{W_b H_b}$
    if $1.1 \leq C_b \leq 1.6 \cap E_b \geq 0.5$ then
        $\mathcal{V} \leftarrow \mathcal{V} \cup CE_b$
    end
end
Align image and crop digits:
$rectRef \leftarrow$ Rectangle((0,0),W,H)
$I^A \leftarrow$ GeometricTransfor(rectRef,$\mathcal{V}$,$I^B$)
$sum1dX \leftarrow \sum_i \{I^A_{i,j}\}$
$sum1dY \leftarrow \sum_j \{I^A_{i,j}\}$
$peakX \leftarrow$ peakDetect($sum1dX$)
$peakY \leftarrow$ peakDetect($sum1dX$)
for $i = 1$ to $|peakX| - 1$ do
    for $j = 1$ to $|peakY| - 1$ do
        $m \leftarrow peakX_i$
        $n \leftarrow peakY_j$
        $m' \leftarrow peakX_{i+1}$
        $n' \leftarrow peakY_{j+1}$
        $d \leftarrow \{I^A_{i,j}\}^{m+m',n+n'}_{i=m,l=n}$
        $\mathcal{D} \leftarrow \mathcal{D} \cup d$
    end
end
Return $\mathcal{D}$
```

detection, we can find image borders and extract the images accordingly.

IV. Legibility Criterion

All datasets were examined under same criteria to evaluate legibility. The following steps illustrate the procedure.

- All extracted images were grouped into ten separate folders corresponding to their digits. Then all images in each folder were examined by at least two people separately. During this stage most of the digits removed were improperly extracted or were blank or contained other numbers.
- The filtered dataset was rearranged into two separate folders corresponding to even or odd digits. They were re-examined for legibility. This ensured minimum priori knowledge was available during examination. Images of different digits that were too similar were ruled out in this stage.
- The entire data was then merged into one single folder and were skimmed one final time to ensure that the data was free of outliers. Some examples of discarded images are shown in Fig. 2



(a) (b) (c)

(d) (e) (f)

Fig. 2. Examples pf handwritten digits discarded during manual checking. 2(a) and 2(b) were eliminated as they were not properly identifiable as a digit five. 2(c) and 2(d) were eliminated for overwriting. 2(e) and 2(f) was discarded as the digits extended out of the bounding box and 2(e) contains part of the grid line

V. Dataset Statistics

After all the illegible digits were pruned, the dataset was divided into training and testing set. The train-test split was done in a 85%-15% ratio for all six datasets. The splitting was done so that all digits in the training set were written by people who did not contribute in the testing set. Also the number of images per digit was kept approximately equal. The statistics of both training and testing sets are given in the Table I.

VI. Tested Models and Benchmarks

We ran several algorithms on the dataset for benchmarking. Due to time and computing limitations, the datasets were tested with lower image sizes. Selecting image size of 28×28 pixels gave accuracy of 54%, and 77% for logistic regression and SVM models respectively. The results jumped to 71% and 84% respectively for image size of 54×54 pixels. We tested on deep learning models as well. On state of the art CapsuleNet model accuracy peaked at 95% at 28×28 pixels.

VII. Conclusion

In this paper, we have assembled data from isolated datasets gathered from over 2500 contributors with a view to maintaining variety within the data. The datasets were checked rigorously following a rigid methodology to ensure legibility of labels. The combined dataset, therefore, has excellent ground truths besides maintaining good variety in terms of age groups,

TABLE I
DATASET SUMMARY

Original Name	Codename	Train-Test Split	Total Digits (Training)	Total Digits (Testing)	Total Digits (Combined)
BHDDB	A	85%-15%	19702	3489	23191
B101DB	B	85%-15%	359	69	428
OngkoDB	C	85%-15%	24298	4381	28679
DUISRT	D	85%-15%	10908	1948	12856
Bangla Lekha Isolated	E	85%-15%	16777	2970	19747
UIUDB	F	0%-100%		495	495
Total			72044	13552	85596

TABLE II
CONTRIBUTOR INFORMATION

Dataset Name	Date of collection	Data Source	Age Range	Male/Female Ratio (%)	No. of contributors	No. of digits per contributor	Frequency of digits per contributor
Bengali Handwritten Digits Database (BHDDB)	17.12.17	BUET, Schools and Colleges in Dhaka	6-20	65/35	260	90	9
BUET101 Database (B101DB)	15.12.17	BUET	18-24	70/30	45	10	1
OngkoDB	17.12.17	Department of CSE, BUET	20-24	70/30	289	100	10
DUISRT	Dec '18	Dhaka	20-24	51.3/48.7	145	90	9
Bangla Lekha-Isolated	Sept'16 to Nov'16	Dhaka and Comilla	6-28	59.4/40.6	1988	10	1
UIUDB	15.12.17	United International University, Mentors'	18-25	75/25	15	40	4

TABLE III
IMAGE COLLECTION AND CURATION INFORMATION

Dataset Name	Medium (Scanner/ Paint)	Formats	Total Number of Digits	Number of Digits Removed	Dimension of images
Bengali Handwritten Digits Database (BHDDB)	Data collected on forms and digitized with a scanner	PNG, 24 BIT COLOR	23400	209	180x180 (Fixed)
BUET101 Database (B101DB)	Data Collected on paper and scanned at 600 dpi	PNG, 24 BIT COLOR	435	7	width: 94 to 110 height: 90 to 110
OngkoDB	Data collected on forms and digitized with a scanner	PNG, 8 BIT GRAY-SCALE	28900	321	180x180 (Fixed)
DUISRT	Scanned from paper forms	PNG 24 BIT COLOR	13133	277	180x180 (Fixed)
Bangla Lekha-Isolated	Scanned from paper forms	PNG, 8 BIT BINARY	20319	572	width: 29 to 267 height: 266 to 180
UIUDB	Scanned from paper, MS Paint, Cellphone Camera	319 JPG, 257 PNG.	576	81	width: 63 to 879 height: 73 to 765

gender and location. The training set can be dowloaded from www.bengali.ai.

ACKNOWLEDGMENT

We are deeply grateful to Bangla Lekha for providing us their dataset. We also must mention the contribution of all the Universities and schools who were kind enough to give their time for our research.