Ekush: A Multipurpose and Multitype Comprehensive Database for Online Off-Line Bangla Handwritten Characters

AKM Shahariar Azad Rabby(✉), Sadeka Haque, Md. Sanzidul Islam, Sheikh Abujar, and Syed Akhter Hossain

Department of Computer Science and Engineering, Daffodil International University, Dhanmondi, Dhaka 1205, Bangladesh
{azad15-5424,sadeka15-5210,sanzidul15-5223,sheikh.cse}@diu.edu.bd, aktarhossain@daffodilvarsity.edu.bd

Abstract. Ekush the largest dataset of handwritten Bangla characters for research on handwritten Bangla character recognition. In recent years Machine learning and deep learning application-based researchers have achieved interest and one of the most significant application is handwritten recognition. Because it has the tremendous application such in Bangla OCR. Also, Bangla writing script is one of the most popular in the world. For that reason, we are introducing a multipurpose comprehensive dataset for Bangla Handwritten Characters. The proposed dataset contains Bangla modifiers, vowels, consonants, compound letters and numerical digits that consists of 367,018 isolated handwritten characters written by 3086 unique writers which were collected within Bangladesh. This dataset can be used for other problems i.e.: gender, age, district base handwritten related research, because the samples were collected include verity of the district, age group and the equal number of male and female. It is intended to fabricate acknowledgment technique for hadn written Bangla characters. This dataset is unreservedly accessible for any sort of scholarly research work. The Ekush dataset is trained and validated with Ekush-Net and indicated attractive acknowledgment precision 97.73% for Ekush dataset, which is up until this point, the best exactness for Bangla character acknowledgment. The Ekush dataset and relevant code can be found at this link: https://github.com/ShahariarRabby/ekush.

Keywords: Bangla handwritten · Data science · Machine learning · Deep learning · Computer vision · Pattern recognition

1 Introduction

There are large numbers of research have been introduced for the handwritten recognition of Latin, Chinese, and Japanese text and characters.On the other

The original version of this chapter was revised: The names of the two Authors have been corrected as "AKM Shahariar Azad Rabby" and "Syed Akhter Hossain". The correction to this chapter is available at https://doi.org/10.1007/978-981-13-9187-3_67

© Springer Nature Singapore Pte Ltd. 2019
K. C. Santosh and R. S. Hegadi (Eds.): RTIP2R 2018, CCIS 1037, pp. 149–158, 2019.
https://doi.org/10.1007/978-981-13-9187-3_14

hand, relatively few research has been done on Bangla Handwritten recognition due to its compliance of Bangla characters and limitation of Bangla datasets. Bangla language has 50 basic characters, 10 modifiers, 10 numerals and more than 300 compound characters. But till now Bangla has no complete dataset that contains all of these characters. So that, recognition of Bangla is at the beginning period contrasted with the techniques for recognition of Latin, Chinese, and Japanese text.

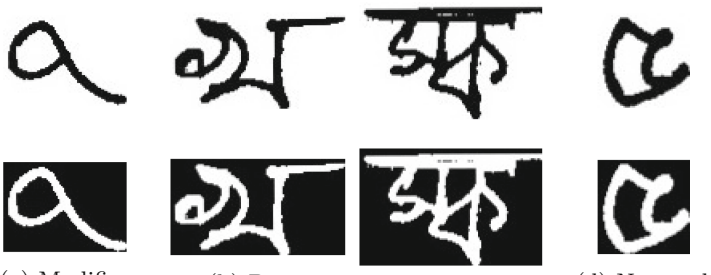
Handwritten character recognition is an imperative issues because of its numerous implementation as Optical Character Recognition (OCR), office robotization, bank check mechanization, postal computerization and also human-PC connections. Bangla Handwritten framework recognition is an extraordinary test since still a long ways behind the human acknowledgment capacity. In this manner, numerous datasets in handwritten recognition area have been accumulated and utilized in different dialects and applications are contrasted with our examination result. There are datasets in CEDAR [1] English words and characters, English sentence dataset IAM [2], Indian [3] for handwritten recognition applications. However, a few studies are attested on handwritten characters of Bangla scripts. Though Bangla is the seventh most spoken language in the world by population and major language in the Indian subcontinent as well as first language in Bangladesh. Though, several studies have been dealing for Bangla handwriting recognition, a robust model for Bangla numerals and Characters classification is still due. One of the main reasons for that the lack of a single comprehensive dataset which covers different types of the Bangla character. There are existing data sets which cover either just the Bangla numerals or just the Bangla characters without modifiers. While it is possible to combine them to form a unified data set, the inconvenience faced by the researchers stems from the lack of consistency in the data presentation of the different datasets. Ekush is the first of a chain of dataset being introduced which aims to Bangla handwriting related research.

In order to support research on recognition for Bangla handwriting, Machine learning, and deep learning applications, we have collected samples (isolated characters) that contain 122-character samples of 4 categories (10 modifiers, 11 vowels and 39 consonants, 52 frequently use compound letter and 10 numeral digit) written by 3086 unique persons include verity of the district, age group and the equal number of male and female. Figure 1 showing the sample of different character images of the Ekush dataset.

2 Literature Review

There are three open access datasets available for Bangla characters, these are the BanglaLekha-Isolated [4], the CMATERdb [5], and the ISI [3]. But every dataset has some drawback. BanglaLekha-Isolated dataset contains 166,105 squared images (while preserving the aspect ratio of the characters), where each sample have 84 different Bangla characters. Those characters have 3 categories which is 10 numerals, 50 basic characters, and 24 compound characters. Two others datasets CMATERdb and ISI where CMATERdb also has 3 categories for

basic characters, numerals and compound characters and ISI dataset has two different categories for basic characters and numerals. A comparison between Ekush and with that three popular sources of Bangla handwriting related datasets (BanglaLekha-Isolated dataset, CMATERdb, and the ISI Handwriting dataset) are given in Table 1.



(a) Modifier (b) Basic (c) Compound (d) Numeral

Fig. 1. Sample images of Ekush dataset

Table 1. Number of images in different datasets

Dataset name	Modifiers	Basic characters	Compound characters	Numeral	Total
ISI	None	30,966	None	23,299	34,256
CMATERdb	None	15,103	42,248	6,000	63,351
BanglaLekha-Isolated	None	98,950	47,407	19,748	166,105
Ekush	30,667	154,824	150,840	30,687	367,018

The Ekush dataset consists of 367,018 images that contain 122 classes and it became the largest dataset for Bangla characters yet. For other language like DEVANAGARI there are some work like "Relative positioning of stroke based clustering: a new approach to on-line handwritten devnagari character recognition" [6] where author used to WCACOM table to collect the data. Same authour has other work like "Character recognition based on non-linear multi-projection profiles measure" [7], They did their experiment on several languages like Roman, Jap-anese, Katakana, Bangla etc. There proposed method used dynamic programming to match the nonlinear multiprojection profile which is used to recognize hand-written characters. "Radon transform" [8] used to produce the nonlinear multi-projection profile. On other paper "Character Recognition based on DTW–Radon" [9] author tries to recognize characters using the "DTW algorithm" [10] at every projecting angle.

3 The Ekush Dataset

Ekush is a dataset of Bangla handwritten characters which can be used as multi-purpose way. Ekush dataset of isolated Bangla handwritten characters structured and organized data was collected from 3086 peoples covering university, school, college students, where approximately 50% 1510 male and 50% 1576 female. The handwriting characters were written in a form after then scanned it to get image data from raw data. We likewise centered around some of different issues for gathering manually written information, such as making a shape, information accumulation strategy, process, programming and pertinent apparatuses. Figure 2 demonstrating a flowchart of making the Ekush dataset.

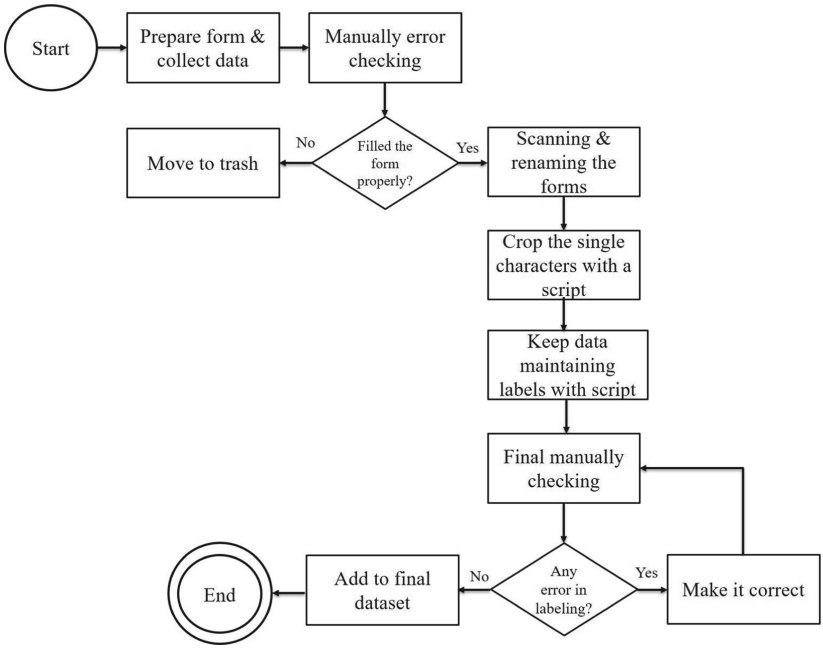


Fig. 2. A flow diagram of "Ekush" dataset

3.1 From Preparation and Collection

Initially creating the form was followed some steps as cell created with equal size, which letter helps to separate the character automatically and made border which help to crop the form equal size. There are lots of compound characters in Bangla language, so we selected most frequently used compound characters.

Than peoples were voluntarily filled up the form which letter scanned for further process.

3.2 From Processing

To process the form we follow the A Universal Way to Collect and Process Handwritten Data for Any Language [11]. During scanning the form, a big black boundary was created to detect the biggest couture. To find that we use canny edge detection algorithm. as same shape with same angle.

Fig. 3. Sample form for Ekush

After skew correction, first cropped the images by row where 12 characters contains each row. After that cropped the row by column. That cropping was

separated all the characters and stored all the images in folder which was labeled by the character name. This process was separate all the 120 characters.

The OTSU [12] algorithm find the best threshold value instead of the specified thresh value. In our method used Gaussian blur for smoothing the images.

We removed all extra bit of information to provide efficient output and stored them. Then Adding white padding to preserve the aspect ratio of the images.

This stored image then inverted into character into white and background as black. All the processed can be done atomically by our designed GUI which able to process 100 scanned images per minute. A filled template used for collection of data is shown In Fig. 3.

3.3 Constructing Ekush

Ekush dataset of Bangla handwritten characters were collected from 3086 peoples where 50% male and 50% female. Each individual developed 120 characters including 50 Bangla basic characters, 50 frequent Bangla compound characters, 10 modifiers, 10 digits. But after processing the dataset we found more than 55 type compound characters because of Bangla compound characters are confusing and some of are similar. So when the writer wrote they made mistake to understand which character actually exist in the form. Then we manually checked and found 2 compound character most of the people were mistake to understand the proper character. So after that, we picked those 2 characters and added in our dataset and finally got 52 types of compound character. These writers were selected from various age, gender, and educational background groups. Figure 4 showing a bar chart of data samples from different ages base on gender.

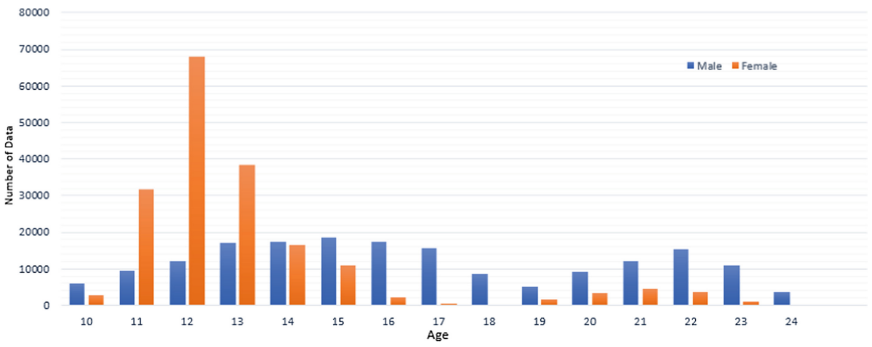


Fig. 4. A bar chart of data samples from different ages base on gender

The collected dataset is preprocessed in three different format:

1. Foreground and background are inverted so that images have a black background with the letter drawn in white.

2. Foreground and background filled with white so that images have a white background with the letter drawn in black
3. Removed noise and smoothing attempted using the Trash holding and Gaussian blur filter. The proceed dataset is further filtered and after necessary smoothing different format are created including CSV.

The Ekush dataset will available on the various format, depending on the user desired applications as well as will be available without extra information on character images, and with padding added to preserve aspect ratio and also in CSV format. Table 2 showing the details of Ekush dataset base on gender.

Table 2. Number of character in Ekush by gender

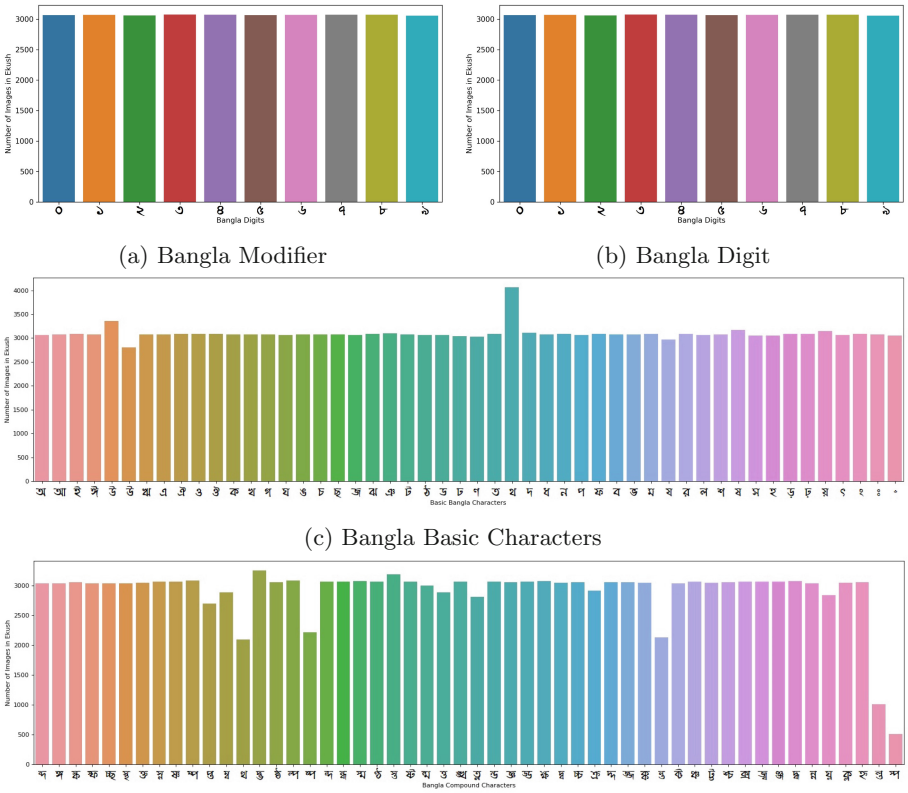
Gender	Modifiers	Basic characters	Compound characters	Numeral	Total	Total in Ekush
Female	15580	78615	76912	15622	186729	367018
Male	15087	76209	73928	15065	180289	

3.4 Visual Representation of the Ekush Dataset

The class breakdown of Ekush dataset shows in Fig. 5, where see that Fig. 5(a), (b) showing Bangla modifiers and digits class as well as seeing that all class are almost equally distributed. The number of the Ekush dataset Modifier and Digits has 30769 and 30687 respectively. From Fig. 5(c), (d) most of the classes are equally distributed, but there are few classes exist where number of images of that characters are not averagely equal to compare the other classes. While writing the compound characters people made mistake to write the proper character. Then we relabeled some compound character also deleted other as well as collected some incorrect compound characters and added them to the dataset in two new class which we are seeing at the last to Fig. 5(d) Compound character bar chart.

4 Data Labeling

The extracted image is saved as the JPEG format and each image has unique ID or name that represent writer gender, hometown, age, education level and serial number. In order to storing the writer information which helps to identify that person using an ID. So that this dataset not only use for handwritten recognition but also help to predict a person gender, age and his or her location as well as it can help investigators focusing more on a certain category of suspects and forensic purposes. This id or name fixed according to the following criteria, Its first one digit indicates writer gender. If it is 0 which means writer was male and



(a) Bangla Modifier

(b) Bangla Digit

(c) Bangla Basic Characters

(d) Bangla Compound Characters

Fig. 5. Visual represent of Ekush dataset

1 means writers were female. After that have writer home district names First 3 or 4 letter, the next one represents age and then their education or occupation level (0 means primary level, 1 high school level, 2 means college level, 3 means university and 4 means other occupation) and the last one is the serial number. And that information are separated by an underscore (_). Here an example.

0_Dha_20_3_00052

Here the first digit one so it was written by a male writer and he is from Dhaka district the next one is 20 which means his age 20 and he is a university student and the last one is the serial number of male data.

5 Possible Uses of Ekush Dataset

This dataset can be used for handwritten character recognition, Bangla OCR, Machine learning and Deep learning base research fields. The prediction of age,

gender, the location from handwriting is a very interesting research field. These information's can also be used for forensic purposes, where it will help investigators focusing more on a certain category of suspects.

6 EkushNet

To test the performance of Ekush dataset we built a multilayer CNN model EkushNet [13] in order to classifying Bangla Handwritten Characters. Figure 6 showing the CNN architect.

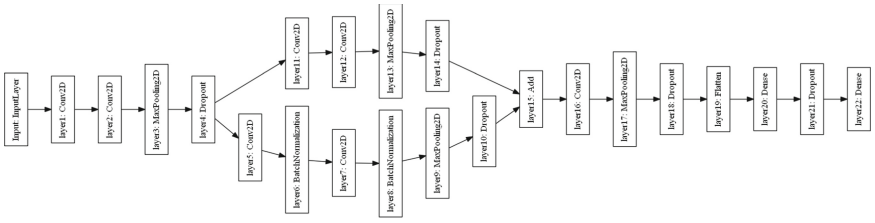


Fig. 6. The architect of EkushNet

We trained the EkushNet and got 96.90% accuracy on training and 97.73% accuracy on validation of Ekush datasets. After train the model we cross validate that with CMATERdb dataset and got satisfactory accuracy of 95.01%. Figure 7 showing the training and validation loss and accuracy on Ekush Dataset.

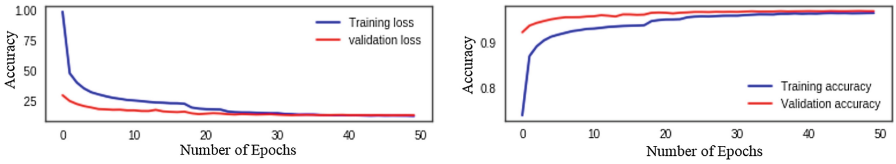


Fig. 7. (a) Training and validation loss. (b) Training and validation accuracy of Ekush

7 Conclusion and Future Work

This research formulate a diverse repository containing primarily for the Computer Vision and NLP research and called the dataset Ekush. In future, we will extend our dataset including all kinds of compound characters. Also, we will make a website where the user can download the form and upload the scan copy which will automatically process those data and added to the dataset after verifying. That website also gives the user to search and download character data by age, gender and districts.

Acknowledgement. I would like to express my deepest appreciation to all those who had provided us the possibility to complete this research under the Daffodil International University. A special gratitude we give to our university and Daffodil International University NLP and Machine Learning Research LAB for their instructions and support. Furthermore, I would also like to acknowledge that, this research partially supported by Notre Dame College, Mirpur Bangla School, Dhanmondi Govt. Girls' High School, Shaheed Bir Uttam Lt. Anwar Girls' College and Adamjee Cantonment Public School who gave permission to collect data from their institution. Any errors are our own and should not tarnish the reputations of these esteemed persons.

References

1. Singh, S., Hewitt, M.: Cursive digit and character recognition in cedar database. In: Proceedings 15th International Conference on Pattern Recognition, ICPR-2000, vol. 2, pp. 569–572 (2000)
2. Marti, U.-V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. Int. J. Doc. Anal. Recogn. **5**(1), 39–46 (2002)
3. Bhattacharya, U., Chaudhuri, B.B.: Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 444–457 (2009)
4. Biswas, M., et al.: Banglalekha-isolated: a multi-purpose comprehensive dataset of handwritten bangla isolated characters. Data in Brief, 12, 103–107 (2017)
5. Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M., Basu, D.K.: Cmaterdb1: a database of unconstrained handwritten Bangla and Bangla-English mixed script document image. Int. J. Doc. Anal. Recogn. (IJDAR) **15**(1), 71–83 (2012)
6. Santosh, K.C., Nattee, C., Lamiroy, B.: Relative positioning of stroke-based clustering: a new approach to online handwritten devanagari character recognition. Int. J. Image Graphics **12**, 1250016 (2012)
7. Santosh, K.C., Wendling, L.: Character recognition based on non-linear multi-projection profiles measure. Front. Comput. Sci. **9**(5), 678–690 (2015)
8. Deans, S.R.: Applications of the Radon Transform. Wiley Interscience Publications, New York (1983)
9. Santosh, K.C.: Character recognition based on dtw-radon. In: 11th International Conference on Document Analysis and Recognition - ICDAR, pp. 264–268, September (2011)
10. Liberman, M., Kruskall, J.B.: The symmetric time warping algorithm: From continuous to discrete. In: Time Warps, String Edits and Macromolecules: The Theory and Practice of String Comparison, pp. 125–161. Addison-Wesley, Boston (1983)
11. Shahariar Azad Rabby, A.K.M., Haque, S., Shahinoor, S.A., Abujar, S., Hossain, S.A.: A universal way to collect and process handwritten data for any language. Procedia Comput. Sci. **143**, 502–509 (2018). 8th International Conference on Advances in Computing & Communications (ICACC-2018)
12. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
13. Shahariar Azad Rabby, A.K.M., Haque, S., Abujar, S., Hossain, S.A.: Ekushnet: using convolutional neural network for bangla handwritten recognition. Procedia Comput. Sci. **143**, 603–610 (2018). 8th International Conference on Advances in Computing & Communications (ICACC-2018)