International Journal of Scientific Research in Computer Science and Engineering
Vol.9, Issue.6, pp.105-109, December (2021)

Research Paper

E-ISSN: 2320-7639

A Paper on Preparation of Dataset for Handwritten Dzongkha Alphabets

Deewas Chamling[1], Yeshi Jamtsho[2*], Yonten Jamtsho[3]

[1,2]College of Science and Technology, Royal University of Bhutan, Rinchending, Bhutan
[3]Gyalpozhing College of Information Technology, Gyalpozhing, Bhutan

[*]*Corresponding Author: yjamtsho.cst@rub.edu.bt, Tel.: +975-17962261*

Available online at: www.isroset.org

Received: 01/Nov/2021, Accepted: 20/Nov/2021, Online: 31/Dec/2021

***Abstract*—** In this paper, we present the complete methodology of preparing a dataset for handwritten Dzongkha alphabets of Bhutan to promote the development of the Handwritten Dzongkha Alphabet Recognition System (HanDARS). The dataset consists of 30 classes, each representing a character of the Dzongkha language with 500 images in each class amounting to a total of 15000 images. The images were manually collected from different individuals and were then augmented to add more varieties to the dataset. The alphabet images were converted to binary format. This dataset can be utilized as a basis for further research and development in the field of optical character recognition for the Dzongkha language. In the future, a greater number of handwritten alphabets needs to be collected to introduce variations in the dataset.

***Keywords*—** Deep Learning; Convolutional Neural Networks; Dzongkha, Bhutanese dataset

I. INTRODUCTION

Dzongkha is the national language of Bhutan that reflects the country's unique culture. The importance of preserving the language whilst promoting it is essential. Optical Character Recognition (OCR) is the process of identifying the character precisely from a printed document, handwritten text or from an image. It is the subfield of computer vision that attempts to find patterns to analyze and understand the text in the image similar to what human visual performs [1].

Bhutan has a huge number of documents that contain precious information about the history and culture of the country. It has become vital to digitize the handwritten Dzongkha text for processing and storing. To accomplish the task of digitizing the handwritten Dzongkha text, it is important to build a handwritten Dzongkha text recognition system. Such systems require a dataset for it to be able to efficiently recognize the alphabets. OCR in various languages have been done which contains a dataset of Devanagari text of India, Farsi text of Arab and so on. However, there has been no advancement in the Dzongkha language. This study presents the complete methodology for the creation of the Dzongkha handwritten alphabets dataset that will help the model to efficiently recognize the alphabet and aid anyone wishing to work on similar fields of OCR and image processing.

In section II, the related works are discussed followed by the methodology on the collection of the handwritten Dzongkha dataset which is elaborated in section III. The results obtained via the preparation and processing of the dataset are discussed in section IV and the summary of the paper with insights into the future work is included in section V.

II. RELATED WORK

Various works have been carried out on Optical Character Recognition and on the preparation of datasets for different languages which has vastly promoted the research and development of image recognition. OCR is the process of recognizing a character and converting it into editable texts, there are six major stages in character recognition namely image acquisition, pre-processing, image segmentation, feature extraction, image classification and post-processing carried out in the mentioned sequence [2].

Modern National Institute of Standards and Technology (MNIST) [3] handwritten dataset contains 60,000 handwritten digit images and 10,000 for testing. It contains 10 classes and is the most popular dataset which has become a standard for various pattern recognition and machine learning algorithms with each image size of 28*28 pixels that has been binarized [4].

ISI (Indian Standards Institute) Kolkata Odia dataset consists of 5790 images of Odia numerals that were collected from 356 individuals via 105 pieces of mail, 166 job application forms and other uniquely developed forms. The dataset was split into 4790 training sets and 1000 testing sets [5].

The authors in [6] created a Bangla dataset consisting of 37,858 images with 50 classes. These images are not uniformly distributed as there contain some alphabets that are rarely used in a proper noun. The training set consists of 500 samples from each class and the remaining for the testing set.

Multi-tech Research Group Online Handwritten Tibetan Character (MRG-OHTC) dataset contains 910 Tibetan classes that were written by 130 persons using an electronic pen on a digital tablet [7]. They split the total dataset of 118,300 into 91,000 training sets and 27,300 testing sets.

III. METHODOLOGY

This section describes the methodology proposed to prepare the dataset. The proposed methodology for the study is given in Figure 1.

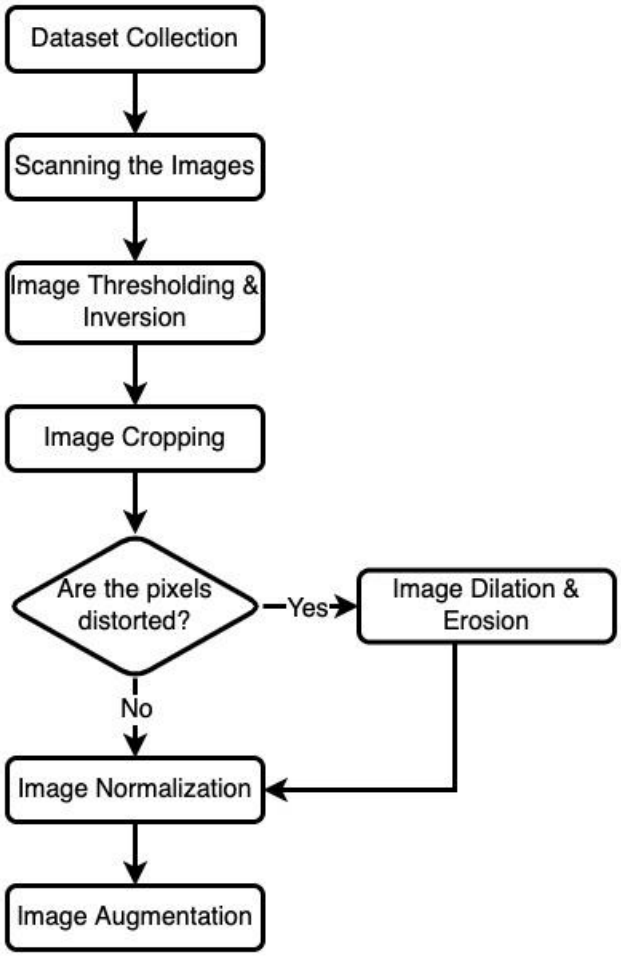

Figure 1. Proposed methodology

Dataset preparation is a crucial process as a good dataset aid in developing a good model. While there are various datasets available online that has been processed and is easy to use, there isn't any related to the Dzongkha alphabet, so a custom dataset was prepared with many processes in between for it to be ready.

A. Dataset Collection

A custom dataset requires the repetitive task of having to go and collect a dataset from everyone in a sheet. Instructions must be made clear so that the people have a clear idea of how and where to write the alphabet. It is important to have spaces in between the letters to allow easier cropping of the letters while processing it. Having an individual space for each alphabet in the sheet greatly reduces the task of constant monitoring of the spaces they maintain in between the consecutive letters. A solution for this approach was using a flip chart. A flip chart was used since it contains numerous boxes that helped in segregating the alphabets in each box. A single paper flip chart contains 165 (11*15) boxes. This was cut into 30 (5*6) boxes for the thirty alphabets to fit. The cut flip chart was handed out to the people willing to write the Dzongkha alphabet. Monitoring them to avoid unnecessary mistakes was one challenge as this took a lot of time for each person. Figure 2 shows the resultant flip chart after being populated with their handwritten alphabets. A total of one hundred data were collected.

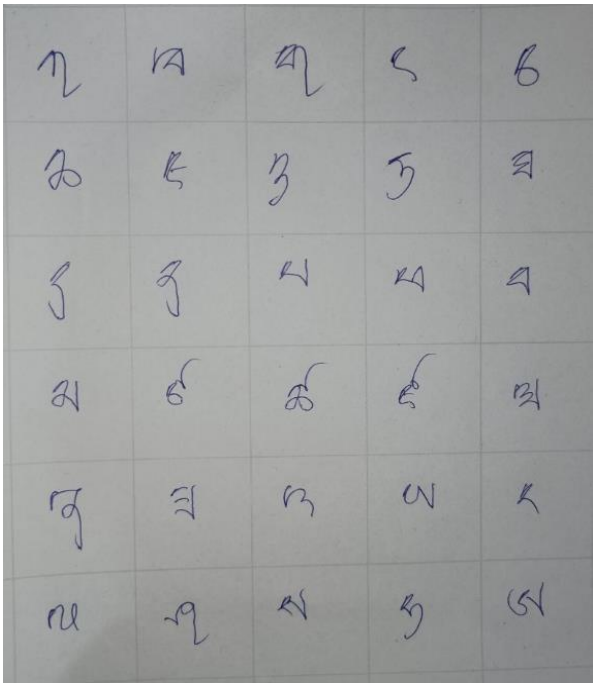

Figure 2. A dataset collected from an individual

B. Scanning the Images

A dataset can be captured via various devices with varying megapixels of a camera at different environmental lightings. This can cause noises in the dataset which is not required. Therefore, scanning is required as this ensures the image to be undisturbed by environmental factors. As the boxes in the flipchart are not required, scanning removes the boxes as well since the lines of boxes are thin. The output from the scanning is shown in Figure 3.

Figure 3. Scanned image of the collected dataset

C. Image Thresholding and Inversion

Thresholding is a method of converting a multi-colored image into a black and white image (two-toned image). This is necessary in order for the model to extract the information and features from the machine more distinctively and easily. As there are different threshold values for different images, finding out the threshold value for each can be time consuming and inefficient. Nobuyuki Otsu developed an efficient algorithm named OTSU's method that performs automatic thresholding to any image [8].

Inversion is the method of reversing the foreground and the background color such as black to white and vice-versa. This is carried out to detect the minute details of an image and process it according to if there are some faults with the dataset. The output from the Otsu's thresholding is shown in Figure 4.

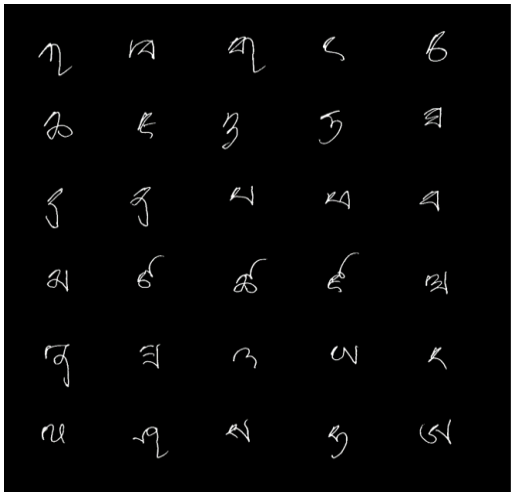

Figure 4. Thresholded Image

D. Image Cropping

After the thresholding and the inversion of the image, each alphabet in the image needs to be cropped in order to segregate into its own label of alphabets which is called class in machine learning. Using the mouse clicking event function available in OpenCV [9], each alphabet was cropped manually for the alphabet. This step is done repetitively for every alphabet. The sample cropping is shown in Figure 5.

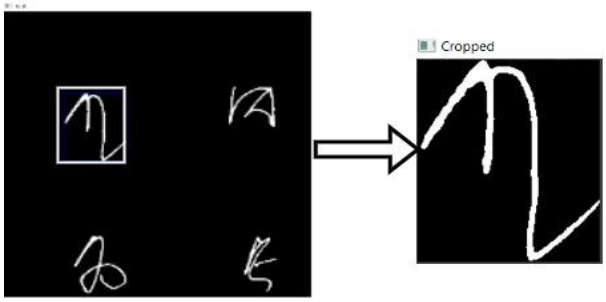

Figure 5. Image cropping. LEFT: Original image. RIGHT: Cropped image

E. Image Dilation and Erosion

While most of the image gets processed correctly, some may lose or gain more pixels in the process of thresholding and inversion. This may cause the model to gain insights into some information that is unnecessary. To rectify these images, morphological processes such as dilation and erosion are used as shown in Figure 6.

Dilation is used when the image has lost some of its pixels. It brings out the features by increasing the object area of the image. Erosion is used when the image has more pixels. It erodes away the boundaries of the image decreasing the object area. In this dataset, there have been no cases where the pixels were more than required, therefore, erosion was not used. However, dilation was used in some cases where the features were diminished and not as required.

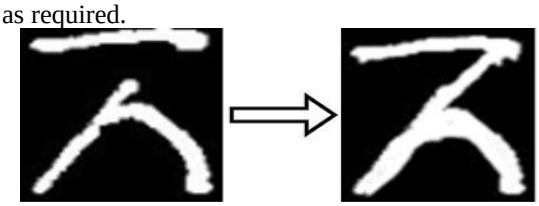

Figure 6. LEFT: Image before dilation. RIGHT: Image after dilation

F. Image Normalization

Image normalization refers to changing the values of the intensity of the pixels in the image. It is mainly used to standardize the image size according to the user with no predefined fixed value. For this project, all the images were normalized into 32*32-pixel images as this were the lowest value the image could be standardized below which the images were not clearly visible (Figure 7).

Figure 7. Images normalized to 32*32 pixels

G. Image Augmentation

Data augmentation is the method to increase the amount of data without collecting new data. Some of the techniques used in augmentation are flipping, rotation, zooming and using various contrasts. Rotation, blurring and varying contrasts were the augmentation techniques used in this project. Five different images were produced using augmentation for each alphabet. Since there was a total of 3000 datasets collected, augmenting five images for each dataset resulted in a total amount of 15000 datasets. The sample handwritten alphabet is shown in Figure 8.

Figure 8. Images obtained after using augmentation techniques

IV. RESULTS AND DISCUSSION

Due to the lack of available dataset on handwritten Dzongkha alphabets, 3000 raw images were manually collected by asking out individuals to write which were then scanned. Thirty classes comprising of 500 images in each class were produced to create a dataset that will be trained on a recognition model to identify the different alphabets present in the Dzongkha language. The total dataset sums up to 15000 images which have been split into two sets with the ratio of 70:30 for the training set and testing set respectively, yielding 10,500 training and 4500 testing images. The sample dataset is shown in Figure 9.

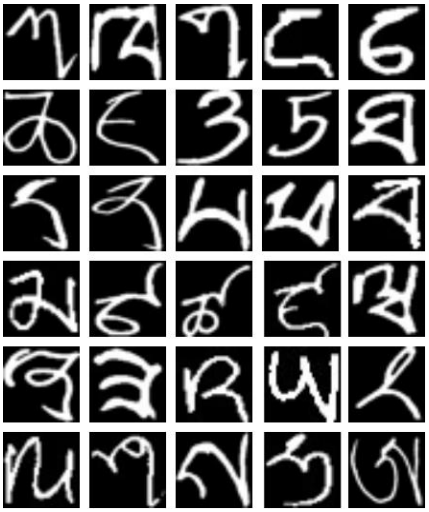

Figure 9. Sample Dzongkha handwritten alphabet

V. CONCLUSION AND FUTURE SCOPE

This paper illustrated the thirty different classes of Handwritten Dzongkha Alphabets Recognition System. 3000 raw images of the dataset were collected manually with 100 images for each class which were augmented to produce image variants resulting in a total of 15000 images. The number of images can be further improved by collecting additional samples from other sources to get an increased number of raw images which will aid the recognition model to identify the alphabets better. This paper lays the groundwork for future studies into handwritten Dzongkha alphabet recognition. In the future, more handwritten alphabets need to be collected to have variation in the dataset. To have a clear pixel in the handwritten alphabets, certain guidelines need to be framed.

REFERENCES

[1] D. Chamling, Y. Jamtsho, and Y. Jamtsho, "Handwritten Dzongkha Alphabet Recognition System using Convolutional Neural Network," *Int. J. Sci. Res. Comput. Sci. Eng.*, Vol. **9**, Issue. **5**, pp. **20–24**, **2021**.

[2] S. K. Dasari, S. Mehta, and D. Steffi D.D, "Optical Character Recognition of Devanagari Script Using Machine Learning- A Survey," *J. Xian Univ. Archit. Technol.*, Vol. **12**, Issue. **8**, pp. **593–599**, **2020**.

[3] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]," *IEEE Signal Process. Mag.*, Vol. **29**, Issue. **6**, pp. **141–142**, **2012**.

[4] W. Zhu, "Classification of MNIST Handwritten Digit Database using Neural Network," *Aust Natl Univ*, p. **7**, **2012**.

[5] U. Bhattacharya and B. B. Chaudhuri, "Databases for research on recognition of handwritten characters of Indian scripts," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. **789–793, 2005**.

[6] 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 Anal Applic*, Vol. **15**, Issue. **4**, pp. **445–458**, **2012**.

[7] L. Ma, H. Liu, and J. Wu, "MRG-OHTC Database for Online Handwritten Tibetan Character Recognition," in *2011 International Conference on Document Analysis and Recognition*, pp. **207–211**, **2011**.

[8] N. Otsu, "A Tlreshold Selection Method from Gray-Level Histograms," pp. **62–66**, **1979**.

[9] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, **2000**.

AUTHORS PROFILE

Deewas Chamling has always been keen on learning more about the world of Information Technology. He completed his Bachelor's degree in Engineering in Information Technology from the College of Science and Technology located in Phuentsholing, Bhutan in the year 2021. His particular interest in Artificial Intelligence led him to choose AI as his preferred elective in his college and completed his final year project successfully in the same field. He wishes to explore more on the various fields of Artificial Intelligence throughout his career journey from his mentors and colleagues.

Yeshi Jamtsho received M. Eng in Computer Engineering from Naresuan University, Thailand in 2020 and a B.Eng in Information Technology from the College of Science and Technology (CST), Royal University of Bhutan

(RUB), in 2014. He has also received a Postgraduate Certificate in Higher Education (PgCHE) from Samtse College of Education, RUB in 2018. Currently, he is working in CST, RUB as an Associate Lecturer since 2014. Deep Learning, machine learning, Artificial Intelligence, Natural Language Processing, Blockchain, and Computer Vision are his current research interest.

Yonten Jamtsho is currently working as an Associate Lecturer at Gyalpozhing College of Information Technology, Royal University of Bhutan. He did his BSc (Hons) in Computer Science (2015) from Sherubtse College, Royal University of Bhutan and M.E. Computer Engineering (2020) from Naresuan University, Thailand. He has also done Post Graduate Certificate in Higher Education (2021) from Samtse College of Education. During his master study, he did his thesis in the field of image processing and computer visions. He has 6 years of teaching experience and 2 years of research experience. His research interests lie in the field of computer visions, image processing and data science.

