Bangla Continuous Handwritten Character and Digit Recognition Using Convolutional Neural Network

BY

Fuad Hasan
ID: 172-15-9901

Shifat Nayme Shuvo
ID: 172-15-9836

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

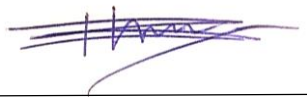
DHAKA, BANGLADESH

MAY 2021

APPROVAL

This Project titled "**Bangla Continuous Handwriting Character and Digit Recognition Using CNN**", submitted by Fuad Hasan, ID No: 172-15-9901 and Shifat Nayme Shuvo, ID No: 172-15-9836 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 1st June, 2021.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Fizar Ahmed — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology

Md. Azizul Hakim — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Sheikh Abujar, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Fuad Hasan

Fuad Hasan
ID: 172-15-9901
Department of CSE
Daffodil International University

Shifat Nayme Shuvo

Shifat Nayme Shuvo
ID: 172-15-9836
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.We really grateful and wish our profound our indebtedness to **Sheikh Abujar**, **Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Computer vision and NLP*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan**, Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

There are few works are available in Bangla Handwritten Character Recognition. To digitalized the analog normal format Handwritten data, we proposed a new methodology to recognized the Bangla character in continuous form. That means in sentence form. We build a system which can take an input of an image of Bangla Handwritten sentence and automatically output the characters exists in the sentence. This system consists of some components like feature extraction, preprocessing, segmentation of character. In Bangla language in written format there is a rigid possibility that two characters are overlapped. This is the main problem in Bangla handwritten format that two consecutive characters overlapped with each other. Some people wrote like this. This becomes difficult to segment the character which overlapped. So, segmentation of the character is important more than to prediction of character with model. To build an OCR system for Bangla Handwritten text, firstly we detect line and segment the line after that word segmentation is done and then individually segment character by character. In this project we used EkushNet dataset model which has the best accuracy till now. This model trained on 85 basic characters, 10 digits, 52 conjunct character, 10 modifiers. By using our algorithm, we can successfully segment 95% true character and recognized by the model. Overall, this current OCR system can deal the recognition system and segmentation of the characters from handwritten Bangla texts effectively.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

CNN	Convolutional Neural Network
NLP	Natural Language Processing
OCR	Optical character recognition
AI	Artificial Intelligence

CHAPTER 1
Introduction

1.1 Introduction

Every region has their own language. Bangla is one of the most popular and speaking language. About 220 million people all over the world uses Bangla language in their daily life. In this present project we work on an OCR system which can recognize Continuous Bangla handwritten language and digits and transform it in a digital information as we can use it in various ways.

OCR is a system which optically read user data which can be language, image, digits in human readable form and convert the data into machine readable format. This type of application system is very popular in first world country. In daily life people use OCR system in Bank, School, College, library automation, post office, reading digital data for blind people and also government documents digitalization.

There are two types of Bangla scripts can be found in our office, home and daily life. One is computer printed character and one is hand written document. The hand written document recognition in sentence level is the most difficult job. In this project we came up with a algorithmic solution.

With the help of NLP and Computer vision this work successfully convert continuous handwritten character to a digital format. The whole work done by few modules like line segmentation, word segmentation, character segmentation and recognize the character with the help of CNN model.

1.2 Motivation

OCR system of Bangla continuous handwritten character can digitalize the analog written text instantly. With the help of an OCR system people can be store data in cloud or server with can be reduce the storage of analog document and can be easily found when needed.

For filing system, people manually input data in the computer which is more time costly and also has some data entry errors. A proper OCR system can do this job with lesser data error, saving the time. This type of system also has an automated sorting, Physical documents can not be tracked by people when need information but digitalized text is flexible when need.

Now data is like gold and information has increased demand day by day. Digitalization of data is necessary in this decade. The huge number of physical data needs huge physical space so we have to keep it in memory with a proper documentation.

Bangla is our mother language. There is not having much work in this field. So, this is a big reason we have to focus on NLP and Computer vision area in Bangla language to build our loving language in digitalized gold.

1.3 Rationale of The Study

Bangla is an old language. It is an Indo-Aryan language and there exists about 230 million native speaker all over the world. Bengali language developed the course more than about 1300 years. This language exists very rich literature and poems. But in the modern technologies not touch this language properly.

In digitalized format for language most of the work can be done by the NLP and Computer Visions. An OCR system can be very helpful to documentation of Bangla language. But handwritten Bangla text is the most difficult job for digitalization. Bangla language is very

difficult to process every character it has 50 characters, 10 modifiers, and much more compound characters. Single character can be recognizing easily but in hand written sentences there is difficulties vary people to people because every person has different way style to write a single character.

In this research we provide a proper solution to avoid the difficulties in continuous hand written sentences. We try to build an algorithm which can preprocess the data properly for the recognition.

The main work in this hole project is building an algorithm to preprocess the sentences. Without segmentation the character an OCR system cannot do the job properly.

1.4 Research Questions

- What is an OCR system?
- How OCR system works in Bangla language?
- What are the benefits of Bangla handwritten character recognition?
- How to prepare the hand written text for recognition?
- How to overcome the difficulties to removing Matra?
- What is the future scope of character segmentation and recognition?
- How recognition model work?
- What is the impact of this research?

1.5 Expected Output

Our research project is totally algorithm base. So, our main target is publishing the paper in a reputed conference. In this area there are other work in Bangla hand written character recognition, but the difficult part not solve by those projects. Most of the work has done same type of input format and variation in the model only. Single character input and recognition by an CNN model.

In this research project the main target is input a continuous Bangla hand written sentence and segment the sentence in word level and character level than recognized by the model. In Bangla language the “Matra” and “Modifiers” detection is the most difficult part of the project there are two types of modifiers upper and lower. We developed a segmentation algorithm which can successfully segment the character level from the sentences.

This OCR system is only for Hand written Bangla character and digits which is recognized by the model and output the digital form in computer display.

1.6 Report Layout

This report contains 6 chapter. Chapter 1 specifying an over view of the project. It also has some sub-sections like 1.1 Introduction of the work, 1.2 Motivation of the work, 1.3 rationale study of this project, 1.4 Research questions, 1.5 Expected output and 1.6 Report layout of the whole research work. In the chapter 2 focused in the Background studies of the work and its sections are 2.1 Preliminaries, 2.2 Related work, 2.3 Research summary, 2.4 Scope of the problem, and 2.5 Challenges. Chapter 3 contains the whole work methodology. In this subsection briefly described the segmentation process how we build the algorithm to solve the difficult work. Sections are 3.1 Introduction, 3.2 Research subject and instrumentation, 3.3 Data collection and preprocessing, 3.4 Statistical analysis, 3.5 Proposed methodology, 3.6 Implementation requirements. In chapter 4 Experimental results is showed and described, this chapter has 4.1 Introduction, 4.2 Experimental results and analysis, 4.3 Discussion. In chapter 5 we discuss about the impact on Society, Environments, and Sustainability. This chapter contain 5.1 Impact on society, 5.2 Ethical expects, 5.3 Sustainability plans. In chapter 6 we declare a conclusion and discuss the future scope of the project. It has 6.1 Summary of the study, 6.2 Conclusion, 6.3 Recommendation, 6.4 Implication of further study. After all of these contain a Reference section which help us some of the related work and study.

CHAPTER 2

Background Studies

2.1 Preliminaries

Bangla continuous hand written character recognition is a system were input a hand written image file and output the information in digital format. To manually digitalized the document is time consuming and also need much more physical space. So, an automated OCR system can improve the time taken and also less error than human. In this chapter we will discuss some related and Background studies which is help us in our project and also similarities and dissimilarities of the work.

Day by day a huge number of government document has to be digitalized to store the data of office and people biodata. So, this is a complete feature extractor for Bangla handwritten text digitalization.

2.2 Related Work

There are some related works which has been done in early year. Now a days, Bangla hand written character recognition is a hot topic to work. Some of the work has done a great job but majority of the work has done the optical character recognition with a single character input. Single character input has no difficulties than continuous character from a sentence. So, our main focus is to segment the sentence in character level and then recognize it with the help of CNN model.

In the age of 1870, the first OCR scanner was introduced by Carey [1] which was a retina scanner. It was works with image transmission system. There are two type of Bangla scripts has found in our daily life. One is printed and another is hand written scripts. In back of previous years there are few works has been done in Bangla printed character recognition system. This works also achieve great success with a good accuracy. For printed character recognition “A complete Bangla OCR system for printed character recognition” [2], An

OCR system for continuous Bengali character recognition [3], “An end-to-end system for web based online handwritten character recognition” [4], “Handwritten character recognition using a hierarchical approach” [5], “Printed OCR system completely describes in this paper” [6] all of these paper work from different author has been completely described their own methodology. Most of the work done in the area of handwritten printed character recognition which is not continuous. From that point of view our work has done a great job with continuous sentences. We extracted all continuous character and all feature are recognized by a CNN model.

2.3 Research Summary

In this research work we developed an algorithm which preprocess the continuous Bangla Handwritten sentences in character level. After that with the help of convolutional neural network we get the predicted output of the character in digital format and display it. For the test purpose we collected 600-700 pages of Handwritten sentences from our university. Then scan those documents and run the algorithm on it. The algorithm works 100% correctly to segment in word level but after word level 95% accurate segmentation done in character level. For recognition purpose we use “EkhushNet” model which has 3lakhs of training data with 97.73% accuracy. It has 50 fundamental characters, all modifiers, 10 digits. There are some errors in few Handwritten character which was falsely recognized but majority of character was successfully classify.

This approach of extracting Handwritten character from a sentence is 95% accurately give us the correct result. Some error happens for the segmentation fault, if two letter seems to be compound letter than this happened.

2.4 Scope of the Problem

Bangla handwritten character recognition is opening a new era in our language. Hand written scripts or story can be easily digitalized by the OCR and CNN. With the help of

our algorithm continuous sentences can be easily segmented to character level and recognized by the model. Computer vision and NLP is connected through one another. Day by day people developed the algorithms and technique to solve different types of problem. In this work our algorithm did not properly work on a connected characters or compound character. This is the only drawbacks of this whole work. In future we will focus on that. The scope of this work will be making a strong algorithm to segment the connected word which is not a character. Different peoples write differently so, it is much difficult to ensure everybody's handwriting will be properly segmented. Thus, this is a new area of research to improve the lacks of current work.

2.5 Challenges

The main challenge was to find out that how we can develop an algorithm such a way in which everybody's handwriting will be covered. But it was not so easy because different people have different type of handwriting, they write their own modified letters, they have different gaps in words, one may use small modifiers, other may use long modifiers. We have to calculate the gap for everybody and find out a global solution which will work for everybody. Bangla language is not like English language, in English they have separate character not connected to each other but in Bangla when we make a word we put "**Matra**" on it. Matra is a single line which connected all characters with the upper bound in Bangla writing format. So, we have to think a way to separate all connected character from a word. This all are big challenges in this whole work to build a such algorithm which can solve all of this challenges.

CHAPTER 3

Research Methodology

3.1 Introduction

In this section, the whole research technique has been described step by step. Here include all of the process and step of the project section. How we preprocess all the data and how we observe the problem and building the solution of the problem. We have used a deep learning model for recognition purpose named “EkushNet”. Which is till now the largest dataset model in Bangla language character recognition.

In this chapter there is subsections which is briefly described the whole methodology. A good methodology is required for further study in this field and for the future work. Work’s flow is given bellow, which is the blue print of the whole work.

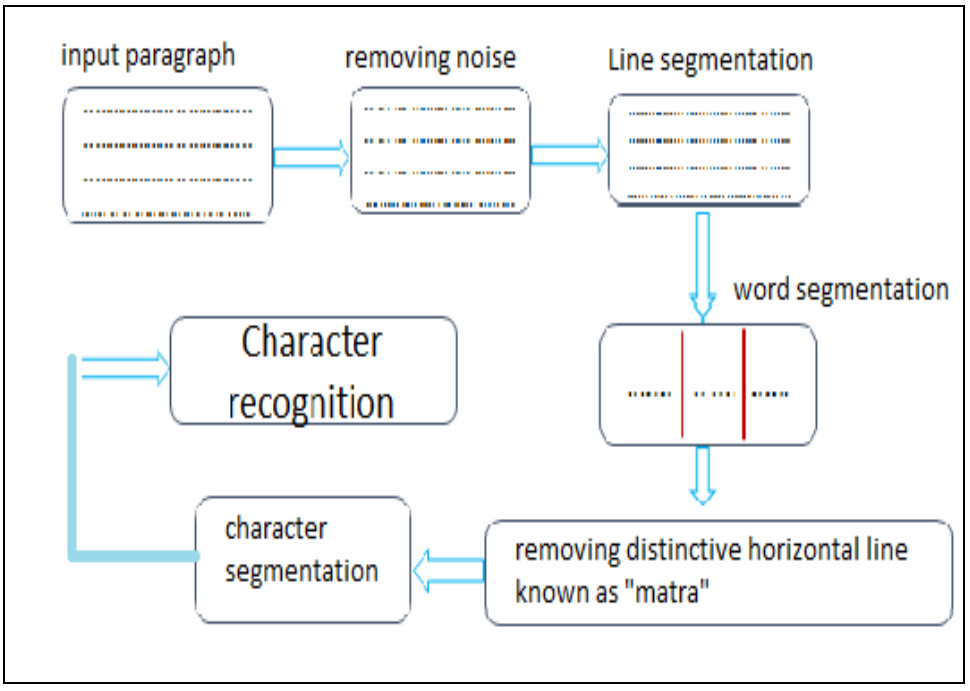


Figure 3.1.1: Illustrates OCR System

3.2 Research Subject and Instrumentation

Our research topic “Bangla continuous Handwritten character and digit recognition using CNN” this is a hot topic in computer vision and NLP. In Bangla language there is less work done in NLP. We discuss the whole process how we build up the algorithm in theoretical and technical process. Deep learning model needs high GPU performance machine and GPU support TensorFlow toolkit. Now the whole list is given bellow, which is used in this project.

Hardware and Software:

- Intel Core i7 7th generation with 8GB RAM
- 1 TB HDD
- Google Colab (12 Gb GPU)

Development Tools:

- Windows 10 or Ubuntu 16.4 LTS
- Python 3.5
- TensorFlow 1.2 GPU engine
- Pandas
- NumPy
- Matplotlib
- OpenCV
- Keras

3.3 Data Collection and Preprocessing

We use our collected data for segmentation purpose. We collected through our university’s student about 600-700 pages of Bangla Handwritten data in a given format. Then

preprocess the data for segmentation to see how our algorithm is work on the data. The necessary steps of all preprocess details are given bellow subsection wise.

The initial step is done by converting the image in gray scale format. Then remove the noise from the data as far as possible. After removing noise convert the image in binary format where all value is 0 or 255. By converting this we found the foreground area of the image.

3.3.a Line Segmentation

In this section we will briefly discuss about the line segment from a passage. Line detection is done measuring the gaps between two words That means, the image is converted into grayscale image. In grayscale image we have only two color's consider character with colour white and the rest of the picture is white. From that we can measure that between two line we gap. Basically, gaps means when we found the summation of the white pixels is zero. As we know the white pixels value is zero. As the image is converted as a 2d matrix. In this image, there some rows that summation is zero. When zero summation row is found it is considered a gap. Line image is cropped according to the gaps given by this method. In recent years many works have been done on line segmentation in different languages English [7], Hindi [8], achieving great success. The image shows in Fig:3.3.1 how segmentation of line is done. In the picture the red colour is indicating the size of line which is considered a line.

Figure 3.3.1: Line segmentation

3.3.b Word Segmentation

In this section we will briefly discuss the word segmentation form the line image. The image is now converted into a binary image. The word segmentation we have done the same work done in the line segmentation. But this time we need to think differently. In this section we found the gaps column wise. When the connected gaps in the column is more than fifteen, we consider it as a gap. Then the gaps are removed from the image. Shown in the Fig:3.3.2. In the image two red lines are considered as a gap. Segmented word shown in Fig:3.3.3.

Figure 3.3.2: Indicating Gaps between two words

Figure 3.3.3: Word segmentation

3.3.c Character Segmentation

In this section we find the segmented word from the word image. In this section the main problem occurs with Bangla language is, its distinctive horizontal line known as “matra". First Problem is to remove the “matra” from the image. Which is connected with all the characters in a word. To solve the problem, we need to resize the main image. As we get the resized image, we crop the upper 25 percent from the image. We can clearly see the “matra” always in the above part of the image. Shown in Fig:3.3.4 with matra and Fig:3.3.5 after matra elimination.

Figure 3.3.4: Before matra elimination

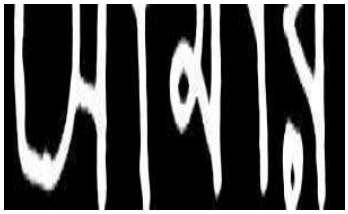


Figure 3.3.5: After matra elimination

Bangla language also has modifiers. There are two parts of modifiers: upper part and lower part. In different types of characters use of the modifiers are different. To detect the modifiers, we have used an algorithm called “Divide and Conquer”. After getting the word image, we divided the image into three parts. To segment the character from the word image, resize the word image into h×w. Three types of word image can be found (i) with no modifier, (ii) upper part or lower part modifier (iii) with both upper- and lower-part modifiers. To identify different types of word with different modifiers there has been a flag indicator set. Shown in Fig:3.3.6.

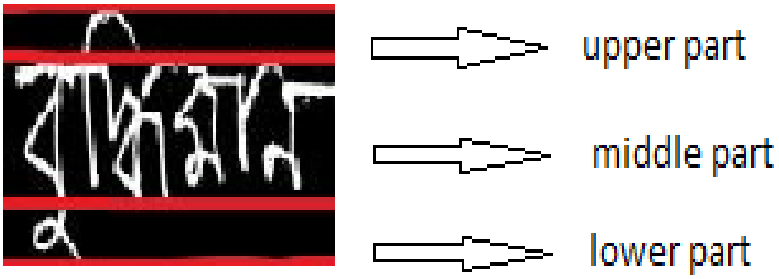


Figure 3.3.6: Word with both upper part and lower part modifiers

3.3.d Word with No Modifiers

Since Bangla language has “matra”, the upper part of the word has been removed from the image. The main body part of the majority of Bangla characters is in the center. As a result, the main body of the word is cropped from the lower portion of the image. Take the main

body component by this equation to exclude matra (Height-25). The height here is 100. The linked white pixels in the y-axis are considered an individual character in that picture, as seen in Fig 3.3.7. After indicating each character and eliminating the unnecessary vertical gaps, each character displayed in Fig. 3.3.8 is eventually separated.

Figure 3.3.7: Indicating Characters

Figure 3.3.8: Segmented characters without modifiers

3.3.e Word with Upper part or Lower part Modifiers

For detection of lower- or upper-part modifier we have to calculate the difference between lower part and the upper part of the image shown in Fig.3.3.10 was taken from the main image in Fig.3.3.9. for labeling the term with single modifiers.

Figure 3.3.9: Upper part modifier detection

Figure 3.3.10.a: Lower part **Figure 3.3.10.b: Upper part**

Figure 3.3.10: Difference between lower and upper part

Then, as seen in Figures 3.3.10(a) and 3.3.10(b), divide the picture into three sections. Each section is 1/3 of the total. If a word has an upper part modifier, at least one-third of the upper part image's black pixels sum will be equal to zero. This is how an upper component modifier has been detected. However, there is no zero (sum of black pixels) in the lower part of the splited image, as seen in Fig 3.3.10(a). In Fig 3.3.11 showed the segmented characters with upper part modifiers.

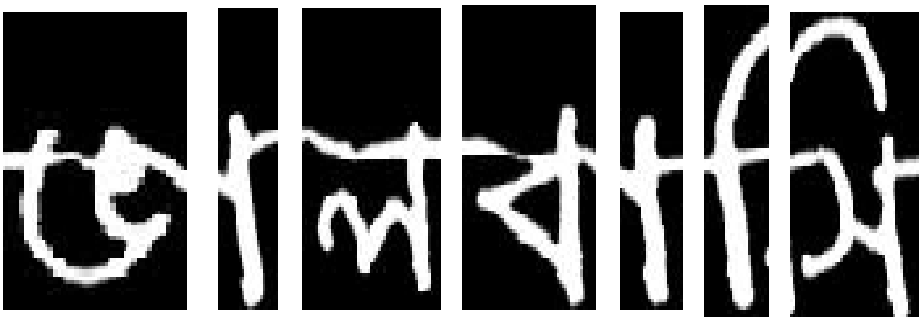
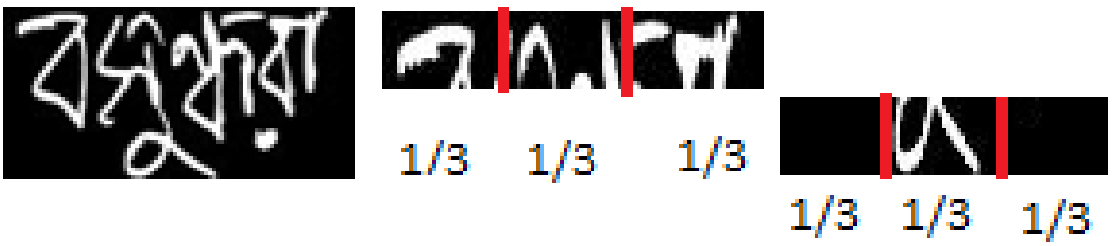


Figure 3.3.11: Segmented character with upper part modifier

If a term has a lower part modifier, at least one element of the representation of the lower component would have zero black pixels. Fig 3.3.12 describes the situation. By marking the word with modifiers, if the word has an upper part modifier, the upper part of the expression is omitted, and if the word has a lower part modifier, the lower part of the image is deleted. Much of the time, the lower part modifier is attached to the word. Calculate the starting white pixel point and ending point vertically from the lower portion of the image to remove the lower part modifier. Since each character has been segmented, the modifier is inserted after the character that belongs to this modifier.



3.3.12.a: Word image **3.3.12.b: Upper part** **3.3.12.c: Lower part**

Figure 3.3.12: Lower part modifier detection

3.3.f Word with Both Upper part and Lower part Modifiers

We already discussed about in the previous two sections, if a word has upper- and lower-part modifiers, at least one subset (1/3) of both the upper and lower parts has at least one zero portion, as seen in Fig 3.3.13. As a result, the word has modifiers for both the upper and lower parts. Then, as described in sections 3.3.d and 3.3.e, the same mechanism would be used for character segmentation. In figure3.3.14 shows that the segmented character with upper- and lower-part modifiers.

Figure 3.3.13: Both upper- and lower-part modifiers detection

Figure 3.3.14: Segmented character with both modifiers

3.3.g Classification and Recognition

The EkushNet [9] model is used here for classification and identification. This model is capable of recognizing characters that we see on a regular basis. It has 50 fundamental Bangla letters, ten digits, and ten modifiers handwritten on it. There are also 50 compound characters in it. The dataset used in this model is Ekush [10]. It is cross-validated using the CMATERdb [11] dataset, with a recognition accuracy of 97.73 percent. Till now this gives us the best accuracy for Bangla Handwritten characters.

3.4 Statical Analysis

In this section we tried to showed that the calculation of truly and falsely recognized character level after segmentation. We collected our data from different people, they have different types of Handwriting. About 600–700-page continuous handwritten data after segmentation likely more then 40-thousand-character recognition done. The truly recognition rate is 95 percent and falsely recognize about 5 percent. In Figure 3.4.1 shows the statical plot of the recognition level.

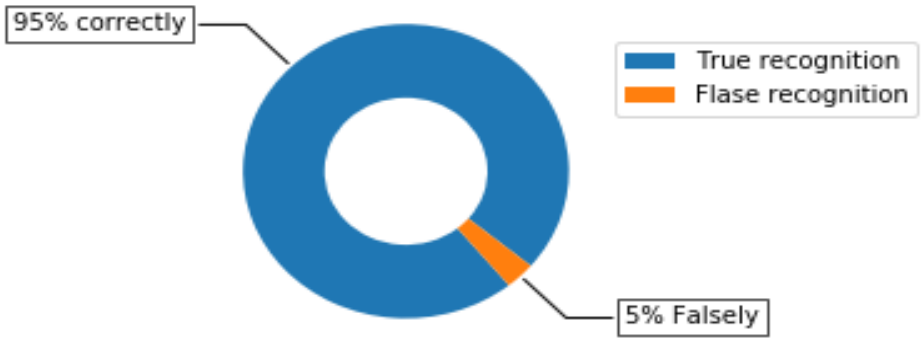


Figure 3.4.1: Plot diagram of statical analysis

3.5 Implementation Requirements

In this section we discussed about the requirements of Library model, for the specific problem. “Bangla Handwritten character recognition” this recognition part will be done by a CNN model. And we need some require analysis to classify the data on the model.

3.5.a Problem Discussion

After successfully segmented all the sentences in character level. We have to classify with a robust model. Here Al has been done a great job to classify all the characters. We use “EkushNet” model which is deep neural network. The main problem was we have to preprocess the segmented data same as the data model has been trained. So, the right segmentation will give us the right recognition.

3.5.b CNN MODEL

The model we use for recognition purpose is called “EkushNet”, this model has 97percent accuracy in test data on Bangla languages. In Figure 3.5.1 shows the model architecture in details.

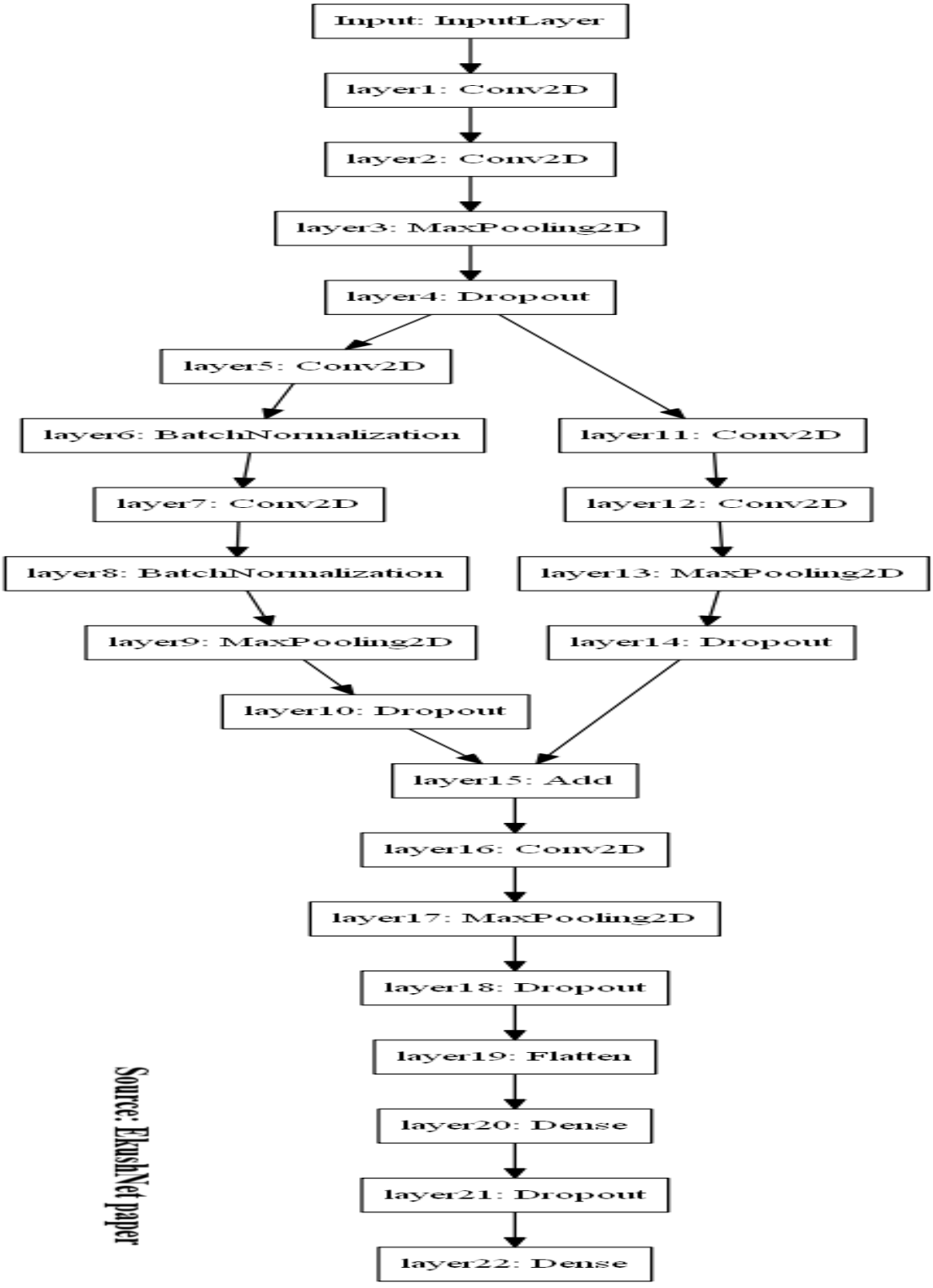


Figure 3.5.1: Architecture of EkushNet

CHAPTER 4

Experimental Results

4.1 Introduction

Character segmentation in Bangla language from continuous sentences is a complex task in computer vision and NLP. We use “EkushNet” for model for the recognition task. In this section we are discuss about the results we get from our collected images after successfully segment the images. A high configuration pc used for train the data for “EkushNet” model. After preprocessing the collected data, we resize it in 28*28-pixel format to each character for recognition purpose.

4.2 Experimental Results

In this section we try to show that how our algorithm works in collected data after segmentation. The machine gives output accurately about 95% character. Some of the character does not recognize properly because of the mismatch of data level or error of classifying the true level. This is more difficult to segment from different people handwriting because every people have different size and shape of the character. If the segmentation works properly then the “EkushNet” model will give 97% accuracy on recognition level. In table 4.2.1 shows that the word image which recognized by “EkushNet” correctly and table 4.2.2 shows that the word image which recognized by “EkushNet” falsely.

Table 4.2.1: Word image recognized correctly

Serial no	Word image	Characters in word image recognized by EkushNet
1		অ া ম া র

2		ব া ং ল া
3		ি ব ে শ ষ
4		ব ক ু ল
5		জ ন ক

Table 4.2.2: Word image recognized falsely

Serial no	Word image	Characters in word image recognized by EkushNet
1		ে স গ ন ে র
2		অ ি প ন া র
3		ব ু ি ধ ম া ন
4		ব য় ু ন্ধ র া

4.3 Descriptive Analysis

To build the algorithm we going through some major steps. Line detection and word segmentation was not so difficult task. After word image to character segmentation was difficult to find the right value to find the gaps between two characters. We take the connected white pixel to consider as a character then cut it from the image. If we found any

black pixel between two white pixels than we consider it as a gap. After cutting all the character we push all in a vector and store that for recognition purpose sequentially. Then load the model with the help of TensorFlow and recognize it. After recognition process we print all the character and modifiers in the display sequentially.

4.4 Summary

This section is for described the output result in short form which is briefly described in experimental results. Our algorithm successfully handles 95 percent of character truly recognized by the model.

CHAPTER 5

Impact on Society

5.1 Impact on Society

The prosperity of technology and computer engineering has formed a remarkable social impact and also global society. There are many major changes impact on society through computer science like home entertainment, helping the disabled, agriculture and also in the medical science. AI done a great job after the 3rd industrial revolution. Now this is the era of 4th industrial revolution between biological, digital and physical combination. AI, robotics, IOT, Quantum computing amalgam towards the 4th revolution. Our AI build algorithm and model also a touch of 4th evolution to accelerate. It will digitalize the nation of Bangla and also other language which is similarity with Bangla. This type of project AI machine completely change the earth we live.

5.2 Ethical Aspects

There are some ethical aspects in research-based project. Our project is well planed and adjusted. We analysis the data properly and all of the information briefly described above is scientifically proven and applied in our own hand. To gain the proper knowledge we have to research properly and maintain plagiarism. All the methodology is described above is our own and the model we use is take permission properly. We also give the independent result of all the output. Therefore, we maintain all the ethical aspects of research that can be take care.

5.3 Sustainability Plans

"Bangla handwritten continuous character and digit recognition" this project will help the community or organization to improve their performance to digitalized their documents. Governments will also use to developed their database of people information. This project

will help financially any type of organization to reduce their worker cost who type the handwritten document for printing purpose. To build capacity of volunteers and trustees, develop marketing and communication plan. Also regularly update the project and fix the bugs. Marketing will attract the donors for funding the project. This will help to build the project more user friendly and compact. This project will survive in the long run because our nation needs this type of software which digitalized the nation firstly. Financial sustainability is the most necessary part of a project this help a project to survive in long run. Community sustainability gives a project a life. If user does not benefit from the project, it will be lost in long run. Our project is not much costly to maintain so, it will survive in the long run.

CHAPTER 6

Conclusion and Future Work

6.1 Summary of the study

This whole project is based on Bangla NLP and computer vision. We have built an algorithm to segmented the continuous sentences in character level. Then recognize the character with the help of a CNN model. This project took us almost 6 months to stand. We done a lot of tasks to give the best possible solution. The steps that we take care was given bellow.

1. Data collection for segmentation purpose
2. Scan all the data and store on drive
3. Extract all the sentences and store on a folder
4. Build an algorithm for segmentation of character
5. Load EkushNet model for recognition
6. After recognition shows the output
7. Check the statical results and then analysis the percentage rate.

This project will help the NLP community of Bangladesh and also impact on the Bangla language to digitalized. This will open a new door towards digitalized nation.

6.2 Conclusion

In this research the main focus was to contribute on Bangla language in NLP and computer vision area. We successfully build an algorithm to segmentation of the character and with the help of AI we recognize it properly. We used an input image of Bangla Handwritten sentences, and the system automatically output the digitalized format of the characters. If two characters are joined together with some pixel value, than we have to differentiate two characters differently. This is another scope of research in Bangla character recognition

and digitalization. This project totally working for normal character and digits with out compound character. Although, this project will increase our NLP recourse in Bangla Language research.

6.3 Recommendation

In future the next step will be improvement the segmentation algorithm. And also try to rebuild the CNN model for more accurate recognition level. And here we are segment with out compound character. In future we will try to segments and recognize the compound character of Bangla Handwritten languages. Few recommendations for Bangla continuous Handwritten character and digit recognition are given bellow:

- Reduce the error rate of recognition
- Segmentation of compound character
- Build CNN model with more training data
- Segmentation of connected two character

6.4 Implications for Further study

We spend a lot of time completing our research work. And many times, our thoughts can't be put into one task so we leave the rest to the future. The future work in a research paper is very important. This is a very important organ. Researchers discuss how much and how they will evaluate and expand future work in the context of their current work. This section contains important research data. And the main thing is that this section gives an indication of the direction or concept of their new research from which those who are new in research will get some direction to start their work. Currently we have identified letters and digits by working with uninterrupted data. Our thoughts on the future are much broader. Since handwritten letter identification is a very difficult task and very few researchers have applied it. So, we're thinking of working with mixed letters and digits. And with the Bengali language there are many letters of the letter that the machine can not recognize easily. We

will work to separate as many characters as there are in that consonant. In our work we have trained models using three lakh data but we are planning to increase the data further in future. So that our model can work more accurately and provide better results.

REFERENCES

[1] J. Mantas, An overview of character recognition methodologies, Pattern Recognition 19, 425-430 (1986).
[2] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[3] Jalal Uddin Mahmud, Mohammed Feroz Raihan and Chowdhury Mofizur Rahman, “A Complete OCR System for Continuous Bangla Characters", IEEE TENCON-2003: Proceedings of the Conferenceon Convergent Technologies for the Asia Pacific, 2003.
[4] S. Bhattacharya, D. S. Maitra, U. Bhattacharya, S. K. Parui, "An end-to-end system for Bangla online handwriting recognition", 15th Int. Conf. on Frontiers in Handwriting Recognition, pp. 373-378, 2016.
[5] S. Basu, N. Das, R. Sarkar, M. Kundu, M. Nasipuri, and D. K. Basu, “A hierarchical approach to recognition of handwritten Bangla characters,” Pattern Recognit., vol. 42, no. 7, pp. 1467–1484, Jul. 2009.
[6] B. B. Chaudhuri, U. Pal, “A complete printed Bangla OCR system,” Pattern Recognition, vol. 31, pp. 531–549, 1998.
[7] G. Louloudisa *, B. Gatosb,1, I. Pratikakisb,1, C. Halatsisa (2009). Text line and word segmentation of handwritten documents.
[8] G. S. Sindhushree, R. Amarnath and P. Nagabhushan (2019), Entropy-Based Approach for Enabling Text Line Segmentation in Handwritten Documents
[9] AKM Shahariar Azad Rabby, Sadeka Haque, Sheikh Abujar, Syed Akhter Hossain, EkushNet: Using Convolutional Neural Network for Bangla Handwritten Recognition, Procedia Computer Science, Volume 143, 2018, Pages 603-610, ISSN 1877-0509.
[10] Ekush: A multipurpose and multitype comprehensive database for Online Off-line BanglaHandwritten Characters, Website: https://github.com/shahariarrabby/Ekush. Last access:20 Jun. 18.
[11] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, “Cmaterdb1: a database of unconstrained handwritten Bangla and Bangla– English mixed script document image,” International Journal on Document Analysis and Recognition (IJDAR), vol. 15, no. 1, pp.71–83, 2012.

Bangla Continuous Handwriting Character and Digit Recognition Using CNN

ORIGINALITY REPORT

10%	4%	8%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Fuad Hasan, Shifat Nayme Shuvo, Sheikh Abujar, Md. Mohibullah, Syed Akhter Hossain. "Chapter 60 Bangla Continuous Handwriting Character and Digit Recognition Using CNN", Springer Science and Business Media LLC, 2020 Publication	7%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		