Islamic University of Technology (IUT)

Department of Computer Science and Engineering (CSE)

Depth Based Bangla Sign Language Data-set Generation

Authors

Ifaz Ahmed Aflan - 154405

and

Junaid Mahmud - 154403

Supervisor

Md. Kamrul Hasan, PhD

Professor, Department of CSE

A thesis submitted to the Department of CSE

in partial fulfillment of the requirements for the degree of

B.Sc. Engineering in CSE

Academic Year: 2018-19

November - 2019

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Ifaz Ahmed Aflan and Junaid Mahmud under the supervision of Dr. Md. Kamrul Hasan, Professor of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Ifaz Ahmed Aflan

Student ID - 154405

Junaid Mahmud

Student ID - 154403

Supervisor:

Md. Kamrul Hasan

Professor

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

Acknowledgement

We would like to express our grateful appreciation for **Dr. Md. Kamrul Hasan**, Professor, Department of Computer Science and Engineering, IUT for being our adviser and mentor. His motivation, suggestions and insights for this research have been invaluable. Without his support and proper guidance this research would never have been possible. His valuable opinion, time and input provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project implementation and finalization which helped us to do our thesis work in proper way. We are really grateful to him.

We are also grateful to **Hasan Mahmud**, Assistant Professor, Department of Computer Science and Engineering, IUT for his valuable inspection and suggestions on our proposal of depth based Bangla sign language data-set generation method.

Abstract

Hearing impaired people have own language called Sign Language but it is difficult for understanding to general people. Sign language is the basic method of communication for deaf people during their everyday of life. Sign digits are also a major part of sign language. So machine translator is necessary to allow them to communicate with general people. For making their language understandable to general people, computer vision based solutions are well known nowadays. In this research work we aims at constructing a model in deep learning approach to generate Bangla Sign Language (BdSL) digits. In this approach there use Depth Based Images to train particular signs with a respective training dataset for acquiring our aim. This model will contribute for moving one step forward to make BdSL machine translator. And open up field for future works.

Keywords: Depth Image, Sign Language, Intel Real Sense, Dataset, Convolution Neural Network (CNN), Hand Gestures.

Contents

1 Introduction

Deaf-mute is a term which was used historically to identify a person who was either deaf using a sign language or both deaf and could not speak [1]. Both are only incapacitate at their hearing or speaking, hence they can do much several things. Communication with the general people which is the only matter that distinct them. The hearing impaired people can simply live like a general person if there is a way for communication between normal people and deaf people. Sign Language is the only way to communication between them. Although hearing impaired people who have sense of sign language, can talk and hear completely. Sign digits are also useful for daily accounting and for communicating the general people and deaf community.

Sign language is a visual language that uses hand shapes, facial expression, gestures and body language [2]. Deaf people share their feeling with various hand shapes and movement in general. A huge amount of research has been done in the field of recognizing Sign Language using different techniques like Hidden Markov Models, skeleton detection, Principal Component analysis (PCA) etc. [3][4][5]. Other great techniques involve fulfill the motion history report associated with gestures, motion capturing gloves and computer vision connected with various colored gloves [6][7].

1.1 Overview

Our main target is to generate the dataset, but for future we will need to incorporate methods to detect different sign languages. So for our approach, there we will use CNN for data classification. A Convolutional Neural Network (CNN, or ConvNet) is a class of deep, feed-forward artificial neural networks that has successfully been applied for analyzing visual imagery[8]. CNNs use relatively little preprocessing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. In our work- first, we speak literature review, then describe our model preparation, then discussion about model optimization and finally the evaluation of model.

1.2 Problem Statement

For BSL, there are some limitations found on the way to build a sign language detection system. There was not a good dataset that could have features. As the previously researches on BSL had an RGB dataset, the edges one of the basic features that could be used. Other than that, the sensor had noise, separation from background was difficult, the color was affected based on different light condition, and etc. problems were found. Based on researches of other sign languages and vision based researches, it is evident that, depth-base image is more suitable to extract geometric features that can be used to formulate different parameters for sign language detection. The problems found in RGB are solved here. Using depth mapping

method, the feature construction and partition will be easier.

1.3 Motivation and Scope of work

About 1.5 million of the total population of Bangladesh, are unable to speak or hear. The only way to communicate with them is sign language. But the sign language is not so common in the population. And also, having an interpreter for the communication is not convenient all the time, and expensive as well. Considering the problem and constraints against it, we thought of creating a system that can translate the sign language into general understandable language in Bengali. So that the communication between the one with inabilities and the general ones can be possible. There has been a lot of research works for recognizing sign languages using different methods in others languages (ASL, CSL etc.), but not much in Bengali. And also, there is not any well-structured dataset for Bengali sign language. The dataset is less reliable with huge training errors and limitations. Therefore, our primary goal is to create a data set with sufficient features that can be further used for detection and translation of Bengali sign language. For that, we will use Intel realsense to capture depth image and depth information so produce the concrete features for further detection. There has been many methods and researches through which it was proven that depth data has more convenience to create features and has less amount of limitations compared to RGB images. There has been a research to detect Bengali Sign Language using CNN [1]. The dataset was created with RGB images and it was claimed to be the first ever dataset created

in Bengali Language in this regard [2]. In Bangladesh there is no open access complete dataset for Bangla Sign characters for research work and development. In this mindset we are thinking to develop an open access dataset of isolated characters for Bangla Sign Language based on depth information. This dataset will help deaf and hearing impaired community by ensuring them to develop education tools. Such as translator, detector of a sign etc. Comparing methods and dataset types of other languages, we can say that, a dataset with depth information and properly extracted features can be appropriate for a suitable method for detection of sign language.

Although different technologies have been developed, But none of the improvement was done to Bangla language gesture recognition. The most recent development and only contribution done to this sector was published September 2018.

To simply put the motivation of our work, we can say-

- Firstly, Lack of dataset in this area.
- Secondly, Only one research has been found that contributes to this area
- RGB images are used for the recent research, hence the output is not absolute for that matter.
- Finally, in outside country we use different detection methods with depth to find greater accuracy.

So, our proposal is to create a depth-based dataset with geometric features of different hand gestures. The data will be created for Bengali sign language alphabets. There are 36 letters (6 vowels and 30 consonants) in BSL. The primary capture will have static images for every letters. Each letter will be considered as a class for detection of sign language. For each gestures, the distance map, palm point, joint angles, finger points and fingers tips will be used to create the Shape of connected joints and internal angles of a specific gesture.

1.4 Research Challenges

The problem with Intel real sense is that it is released very recently. This is why it lacks a lot of library which other devices are with rich with. The problem like this creates obstacles that are solved using various methods. We will be discussing them shortly. The Intel real sense with python version 3.7 doesn't work. Python 3.7 is the latest. Hence why we needed to use either 2.6 or 3.6, both of them lacks few key functions that help us generate more useful output. However this is not the main issue. Due to being very new device, Intel real sense lacks the SDK support and library support. The library provided by Intel Pyrealsense2 doesn't have any functions that would provide output of the hand. So we had to find alternative way to train model to detect the points. Although real sense SDK has great UI, the github repository provided by Intel doesn't contain many examples regarding this matter. Hence why we had to use depth image with another model to train the model. Beside this, the model training required a very high configuration computer. We

had to borrow parts from here and there. Another problem we faced during buying the real sense. We had to buy it from USA and we had to wait for a long time, beside the wait time we had no traces in this sector. Since BdSL is relatively new field with one paper published in this sector, we couldn't find many papers in this sector. The only paper worth mentioning is Isharalipi. Which was published back in September 2018. So we had to research and study many different things to decide on how to approach the research. With depth a lot of problem in RGB model is fixed.

1.5 Thesis Outline

In Chapter 1 we have discussed our study in a precise and concise manner. Chapter 2 deals with the necessary literature review for our study and there development so far. In Chapter 3 we have stated the skeleton of our proposed method, proposed algorithm and also the flowchart to provide a detail insight of the working procedure of our proposed method using Depth based image. Chapter 4 shows the results and comparative analysis of successful implementation of our proposed method. The final segment of this study contains all the references and credits used.

2 Literature Review

Keeping the prominent training loss in consideration, we searched and further found some research works where depth images were used along with different methods for different researches.

2.1 Sign Language

This research goals to construct a model that will identify numbers of BdSL. For recognizing various sign multiple approaches have been used by several researchers which were accomplished in different area.A New Approach of Sign Language Recognition System for Bilingual Users [9] can recognize 11 Bengali digits and 16 words. There they proposed an universal interpreter software for skin detection feature extraction. Theirsystem using a database of (27x10x20) images.Numbers have been recognized effectively in Indian Sign Language Recognition [10]. They represented a framework for a HCI capable of recognizing signs from Indian sign language with PCA (Principle Component Analysis).In Sign Language Recognition using Microsoft Kinect [11] paper, they used computer vision algorithms and build a characteristics depth and motion profile for each sign language digits 0-9. The feature matrix they generated was trained with SVM classifier. But this approach has a dependency on specific camera device.Fine Hand Segmentation using Convolutional Neural Networks [12] proposed a method for recognition very accurate hands gesture views based on Deep Learning architecture. In their model they mapped convolution layers directly to a segmentation

mask with a fully connected layer. They tried to implement it as efficient in real time as possible.A recent work was done for recognizing Nigeria indigenous sign language. There they introduced an Yoruba Sign Language recognition system [13] using image processing and Artificial Neural Networks (ANN). [**?**]

2.2 American Sign Language Alphabet Detection Using Microsoft Kinect

In this method, images were taken in two ways. In total 24 letters were detected, because 'J' and 'Z' were dynamic.first, wearing a colored gloves where different regions were colored differently. Secondly, depth image were taken. For feature extraction DAS (Distance Adaptive Scheme) was used. First, the hand was detected as the closest object to the kinect. A variable threshold value was used that was used to find the end point of the hand. When the endpoint of the hand was found, everything else in any other distance were omitted. In this method, three different datasets were used. A random forest classifier was used to detect the sign language. For validation, 'half and half(h-h)f' and 'last one out(l-o-o)' were used. The accuracy result were different for different papers.

2.3 Recognition of Symbolic Gestures Using Depth Information

The images that are captured are binary at the beginning. Those images do not have much information. The region of interest has to

be separated from the background. So, to extract the features, there have to be some pre-processing. The SIFT based feature extraction method is followed here. After that, the gesture is recognized using SVM. In the proposed methodology, the hand segmentation is done after finding the closest point to the Kinect. Then using a empirical threshold value, the segmentation threshold value. After that, the 'Depth Silhouettes' using depth map are generated using a specific mathematical formula. After finding the points in the data set for a gesture, K-means clustering is used for classification. Using SIFT features, the accuracy of recognition was around 95% and it was found by F-score calculation.

2.4 American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion

In deep learning, basically for CNN a huge number of data is needed to be put in the network which requires a lot of time. To reduce the time, data augmentation is applied that synthesized additional data derived from the original one. Also, it creates more variations in the data because of rotation, scaling etc, by keeping the recognizable features intact. In this method, the 3D cloud of a point is captured using a camera and a number of virtual cameras. Yaw-pitch-roll rotation around the volume center is implemented for each axis. The CNN model is created with strong regularization tech-

nique. Because of inter-class similarities, some gestures are found almost same for the model. For this, multi view inference fusion strategy was proposed in order to augment the speculation of each individual view. The pre-processing part included find the hand portion in the whole image, which covered a small portion of the image. The palm point was calculated using the mass center calculation of depth image and is segmented as a circular region. Mean accuracy of the whole process in different parameter was around 88% . For half half validation method, the accuracy was 99.9% which was found using around 32831 testing dataset. So, this method was useful when there is a huge amount of data.

2.5 A Potent Model to Recognize Bangla Sign Language Digits Using Convolutional Neural Network

Hearing impaired people have own language called Sign Language but it is difficult for understanding to general people. Sign language is the basic method of communication for deaf people during their everyday of life. Sign digits are also a major part of sign language. So machine translator is necessary to allow them to communicate with general people. For making their language understandable to general people, computer vision based solutions are well known nowadays. In this research work we aims at constructing a model in deep learning approach to recognize Bangla Sign Language (BdSL) digits. In this approach there used Convolutional Neural Network (CNN) to

train particular signs with a respective training dataset (Eshara-Lipi) for acquiring their aim. The model trained and tested with respectively 860 training images and 215 (20%) test images of tent classes of digits. Finally,the training model gained about 95% accuracy at recognition of Bangla sign language digits. This model will contribute for moving one step forward to make BdSL machine translator. Now since we have developed everything based on this paper, by providing contribution on the depth images. We are going to discuss in details about this paper and it's methodologies.

Proposed methodology

A neural net is used in this system to recognize hand signs which is Convolutional Neural Network. The neural net layer explanation, dataset properties, data process, model training and many other methodology is discussed.

The Eshara-Lipi dataset which was collected for this project we used to train the model. Eshara-Lipi dataset contains Bangla Sign Language digits from 0 to 9 (0, 1, 2 . . . 9). The dataset has following properties-

- Every class has 100 different images of different peoples hand
- Ishara-Lipi Dataset has total 1000 (10 * 100=1000) images
- All sign images is cropped and resized by 128 x 128 pixels
- Dataset images is formatted in.JPG format
- Images are gray scale and binary coloured then did some preprocessing works

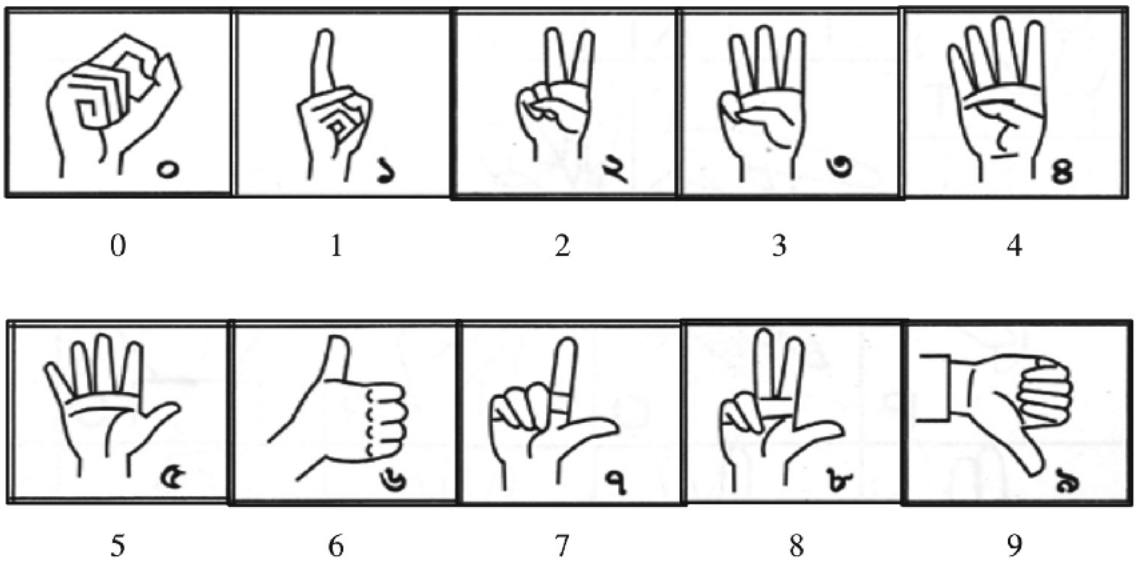

Figure 1: Sign Language

Data Preprocessing

The Eshara-Lipi dataset provides 128 x 128 pixels gray scale images. Some pre-processing works were done for making it usable to train model. Firstly all images were resized by 28 x 28 pixel size. The images were converted into gray scale, then binary coloured image and given the correct labels. Finally saved the image pixels into a CSV file to reduce needed computation power. The method we used determines the threshold automatically from the image using Otsu's method.

Designing The Model

To recognize these digits here used multi-layer convolutional neural networks which are connected each other. The model is represented by multi layered CNN with two sub layers. First two layers are same, there have two convolution layers with same padding and swish (3)

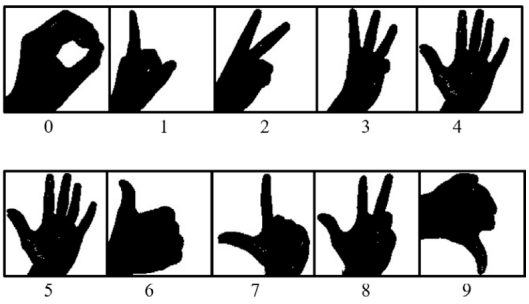

Figure 2: Preprocessed Data

activation using 32 filters and 5:5 kernel. Then also added a max-pooling layer there. The max-pooling layer has 2x2 followed by 25% dropout layer. All dropout layers used here is for reducing over-fitting. The model also use ADAM optimizer [14]. The previous two conv layers generate output and then the output from this two layers goes as an input of two sublayers. The both sublayers contain same 2 convolutional layers with the same swish activation, padding and 64 filters with a 5x5 kernel, followed by another convolutional layer with a 3x3 kernel. The output of last 2 sub convolutional layers added together and go through a Max-Pooling layer. This Max-pool has 20% dropout [15] layer.Then flatten the layers and used a fully connected dense layer with 2048 hidden nodes. Final output layer has 50 nodes with SoftMax (1) activation. Using softmax activation means playing with the logistic regression on the feature extraction before the finally connected layer. In this stage of model, a flatten function is used for shape optimization. The basic concept of applying flatten and dense layer function and its output pattern is shown below (Fig.4 and Fig.5).

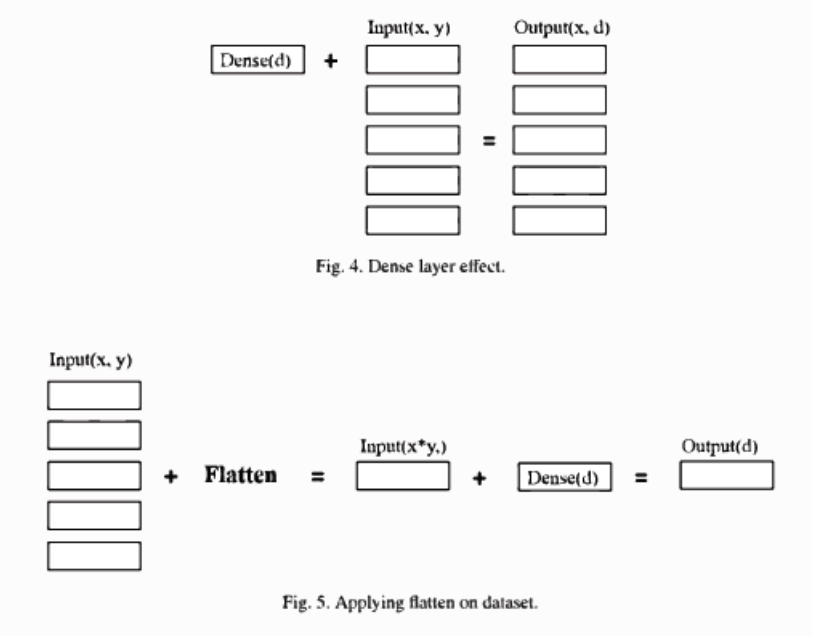

Figure 3: Designing the model

Activation

Nowadays the most commonly used activation function is ReLU (2) by default. The ReLU function is defined by equation is

ReLu(x) = Max(0, x) (2)

The ReLU activation assigns the parameter back to itself. It creates the problem of "dead neurons". There has some better proposed alternative, such as the ELU, SELU and others. Another activation function is used nowadays for efficiency is named Swish activation (3). It's very simple in equation

S wish(x) = x divided by sigma(x)

Model Optimization

Model optimization is used for making the model more efficient and reliable to input data. In this deep learning model here also applied

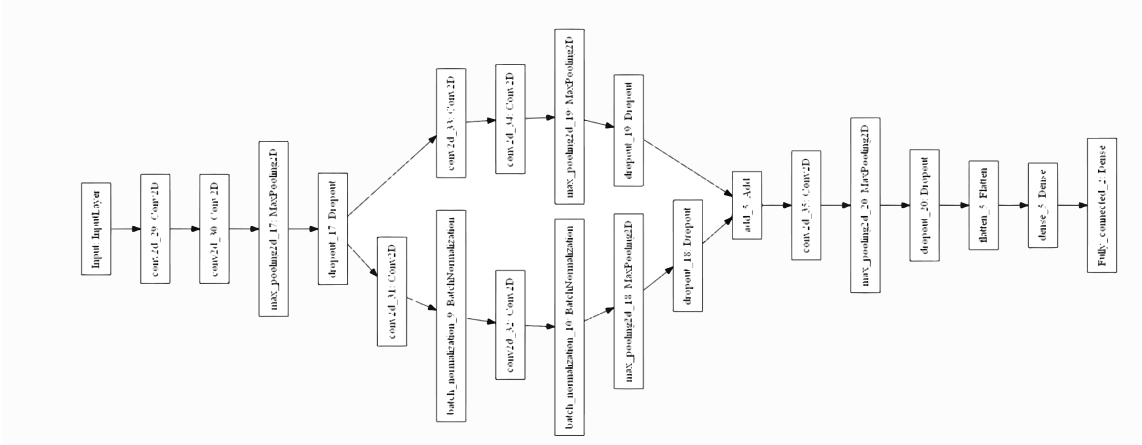

Figure 4: Convolution neural network

some optimization techniques. Used SGD for compiling model as optimizer. Stochastic Gradient Descent (sgd) performs as a parameter for each training example. It is a much faster technique. It usually performs single update at a single time. The cross-entropy is a better choice for cost function optimization. It is known as cross-entropy cost function; also called regularization method. For making better classification and prediction in neural network this function is used widely. Here used a categorical cross entropy as loss function (4).

Model Evaluation

The model developed with Ishara-Lipi dataset performed 94.88% validation accuracy and 95.35% training accuracy. As occurred the training loss and validation loss is shown below in table and graphical representation.

Data Analysis

There has been a lot of research works for recognizing sign languages using different methods in others languages (ASL, CSL etc.), but not

Evaluation	Rate
Training Loss	12.38%
Validation Loss	26.13%
Training Accuracy	95.35%
Validation Accuracy	94.88%

Figure 5: Convolution neural network

much in Bengali. And also, there is not any well-structured dataset for Bengali sign language. The dataset is less reliable with huge training errors and limitations. Therefore, our primary goal is to create a data set with sufficient features that can be further used for detection and translation of Bengali sign language. For that, we will use intel real sense D435 to capture depth image and depth information so produce the concrete features for further detection. There has been many methods and researches through which it was proven that depth data has more convenience to create features and has less amount of limitations compared to RGB images. There has been a research to detect Bengali Sign Language using CNN [1]. The dataset was created with RGB images and it was claimed to be the first ever dataset created in Bengali Language in this regard [2]. In Bangladesh there is no open access complete dataset for Bangla Sign characters for research work and development. In this mindset we are thinking to develop an open access dataset of isolated characters for Bangla Sign Language based on depth information. This dataset will help deaf and hearing impaired community by ensuring them to develop education tools. Such as translator, detector of a sign etc. Comparing methods and dataset types of other languages, we can

say that, a dataset with depth information and properly extracted features can be appropriate for a suitable method for detection of sign language.

2.6 Problem Description

For BSL, there are some limitations found on the way to build a sign language detection system. There was not a good dataset that could have features. As the previously researches on BSL had an RGB dataset, the edges one of the basic features that could be used. Other than that, the sensor had noise, separation from background was difficult, the color was affected based on different light condition, and etc. problems were found.

2.6.1 Problem Identification

Based on researches of other sign languages and vision based researches, it is evident that, depth-base image is more suitable to extract geometric features that can be used to formulate different parameters for sign language detection. The problems found in RGB are solved here. Using depth mapping method, the feature construction and partition will be easier.

A good clustering has minimum intra-cluster distance and maximum inter-cluster distance. This way a clustering performance evaluation can be achieved which will give us insights on how well the data are clustered. A cluster is therefore a collection of objects which are "similar" among them and are "dissimilar" to the objects belonging to other clusters. We can show this with a simple graphical example. This paper represent a deep learning based Bengali Sign Language Digit Recognition System. For sign recognition methods, vision-based models and digit identification methods, convolutional

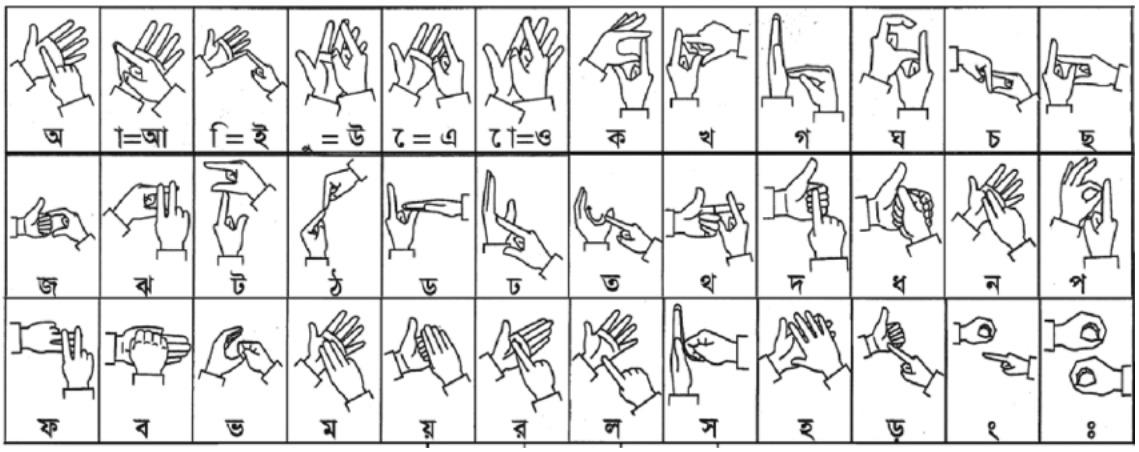

Figure 6: Bangla Sign Language

neural network proves a strong candidature. The proposed models deliver output in text form which support to remove the communication interruption between hearing impaired and general people. For standardization of the Bangla Sign Language, we want to use our dataset and the model as a platform. However, everyone cant understand sign language, in future we will change conversation to sign for pleasant communication between different users. In Future we will reach our database recognize more characters, even to recognize gesture of the Bangla Sign Language and to convert them to Bangla text.

2.6.2 Problem Solution

So, our proposal is to create a depth-based dataset with geometric features of different hand gestures. The data will be created for Bengali sign language alphabets. There are 36 letters (6 vowels and 30 consonants) in BSL. The primary capture will have static images for

every letters. Each letter will be considered as a class for detection of sign language. For each gestures, the distance map, palm point, joint angles, finger points and fingers tips will be used to create the

3 Proposed Approach

Previously, a research related to bengali sign language was found, where the dataset was named as "Ishara-lipi". In Ishara-Lipi dataset, after discarding errors and preprocessing, total 1800 images were created. Characters dataset contained 50 sets of 36 Bangla basic sign characters, collected by the help of different deaf and general volunteers from different institutes. In Bangla Sign Language sign characters there have 6 vowels and 30 consonants by which they can finger spell all Bangla words. This dataset could be used to develop computer vision based or any kind of system that approves users to search the meaning of BSL signs. There were 1000 images for digits, 300 images for vowels and 1500 images for consonants. All the images were taken with a digital camera and so the images were RGB. Then the images were converted to grayscale image and binary images. Then feature extraction filters were applied and hence, the images were put into the CNN-feed forward initial layer. The CNN had two sublayers, first layer had same convolution layers. They were used as inputs for the next layers. ReLU (Rectified Linear Unit) was used as the activation function. There were 12.38% training loss and 94.88% validation accuracy and 95.35% training accuracy.

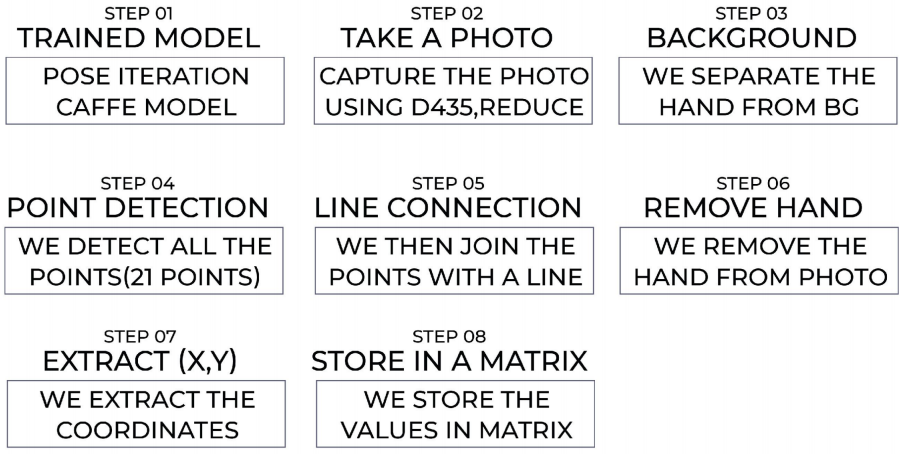

Figure 7: Step by step dataset generation

Following our objective, we will create the data-set. First, we will capture the gestures using a kinect. Following, the palm point will be recognized as it is the center point of the hand. And hand is the closest object to the kinect. Then we will omit the background. Then we will use gray scale mapping of the depth image based on depth. Furthermore, the finger points and finger tips will be identified using Kinect, and joint mapping will be done by normalization. As we have the coordinates, we can measure the joint angles. Then we will merge the features and create the dataset. Joint angle and finger tip positions are important features that can be used to detect sign language.

3.1 Proposed method

Isharalipi used a very good method for detection of gestures using convolution neural networks. However the convolution neural network with RGB doesn't provoide as strong outcome as convolution neural network with depth images.

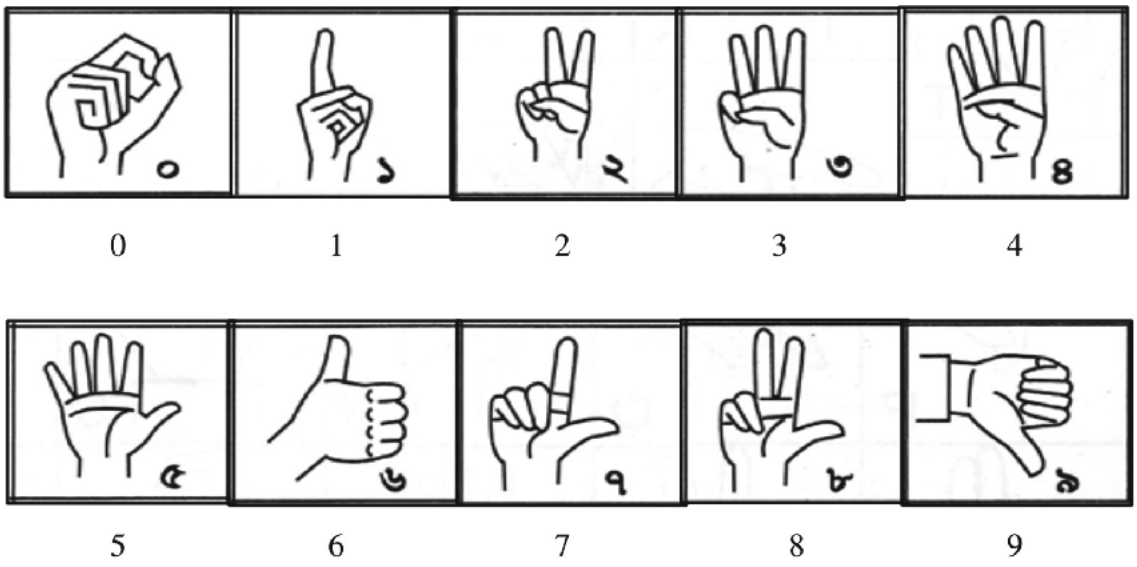

Figure 8: Numbers

So from the above chart it is clear that the accuracy with depth is far superior than any other methodology for convolution neural networks. So for this reason, our target is train a model using depth based image so that it can identify gestures more accurately. So our goal is provide a solution using depth image, using the device intel real sense D435. However obstacle appeared as Intel real sense doesn't have any library function to provide any information regarding hand detection. So we used a already exisiting depth learning model, named post iteration caffe model. This caffe model uses deep learning with depth image to understand the image and detect it's 21 points.

We train a model using depth image and then we detect all the palm points(21 Points) and store the coordinates. Then we use skeleton joining points, to recognize hand signs.

METHOD	METHOD	BRIEF	DATASET
ISHARA-LIPI [1]	TA - 95.35 VA - 94.38 TL – 12.38%	RGB DATA, CNN	28X28 BINARY IMAGE FROM RGB IMAGE
GABOR FILTER [2]	75%	SIFT BASE	DEPTH DATASET STATIC IMAGES
SHOTTON [2]	93.96%	DEPTH, SEMINAL	DEPTH DATASET STATIC IMAGES
KINECT [2]	90%	DAS & EDS	DEPTH DATASET STATIC IMAGES
CNN [3] (KINECT)	98.37%	DEPTH BASE	DEPTH IMAGE SKELETON LINE

Figure 9: Accuracy Table

3.1.1 Dataset Generation

Our target is provide a rich dataset, which can be used for detecting gestures more accurately and precisely using CNN. Our contribution is in the dataset generation. Which provides the output we desire

- Step 1, Firsly we train a model, using post iteration caffee model using the depth images that we have captured through intel real sense D435

- Step 2, the second step of the system is to capture a image using a normal web camera, phone camera or any camera for that matter.

- Step 3, We then separate the hand from the background so that we can detect the hand points more easily. Since we are capturing the images and converting them to binary it's more accurate to detect the hand points.

- Step 4, we then detect the points in the hands using our post iteration caffe model, and it will detect all the 21 points from

the picture.

- Step 5, we then start joining the neighbour points to each other. This way a skeleton joint model is shown.
- Step 6, we then remove all the other objects such as hands, fingers as since we have finally generated the skeleton model.
- Step 7, we extract the Global and Hand boths' (x,y) coordinates.
- Step 8, finally we store the (x,y) values in the matrix and get our features extracted. Hence we generate our desired data sets.

our proposal is to create a depth-based dataset with geometric features of different hand gestures. The data will be created for Bengali sign language alphabets. There are 36 letters (6 vowels and 30 consonants) in BSL. The primary capture will have static images for every letters. Each letter will be considered as a class for detection of sign language. For each gestures, the distance map, palm point, joint angles, finger points and fingers tips will be used to create the Shape of connected joints and internal angles of a specific gesture.

Our target was to implement all the steps within the time frame, however due to lack of knowledge and technical expertise, things didn't go as smoothly. Firstly the obstacle of intel real sense really

delayed our workflow. We weren't able to get anything from it, The pyrealsense2 library doesn't provide us. So we had to find alternative to the real sense, while making sure our target goal is achieved. We although couldn't complete the entire thing, however the output we have generated seems to be nearly done state.

At first we had to make sure that our skeleton model for dataset collection is perfectly made. We one main folder, under the main folder we had our 3 sub folders, namely numbers,letters and vowels. Under number we had 10 more folder and under letters we will have more folder and for vowels we will have 11 more folders. Under each folder there are 10 or more sub-folders where they are named as person1, person2, person3 etc. Then we had to store the pngs under those sub-folders.

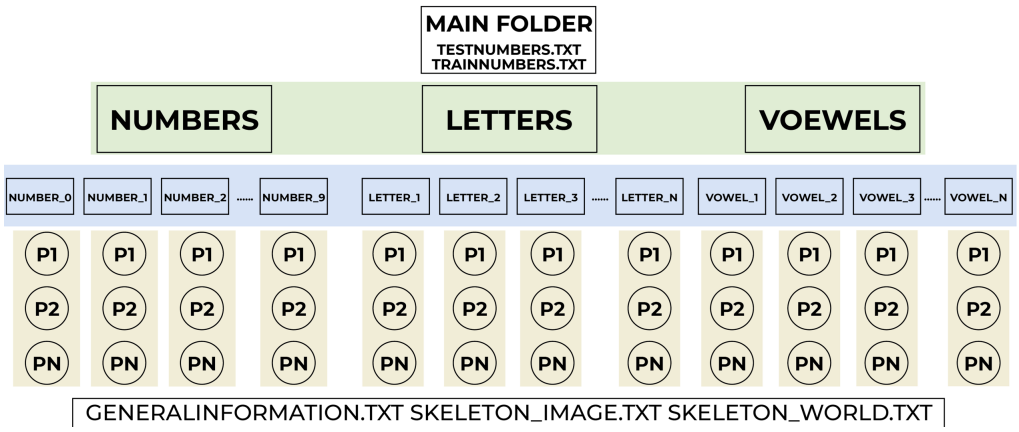

Figure 10: Storing data in hierarchy

Trained model Now after storing the data, we have generate the two files, one is training.txt and another one is test.txt. Training.txt are the classified data that we have already got from trainning the model. Using those training model and comparing the distance be-

tween the points we compare with the test data. The test data then are store.

- Firstly after training the model we take a normal picture of our hand with background via any camera. Below a figure is given.

Figure 11: Capturing a normal image

- Detect the points in the hand, a total of 21 points is detected from the picture.

- We then separate the entire image and only keep the red balls that we have detected in the image with black background.

- After that we join all the points from one point the the neighbouring points to create a skeleton model.

Figure 12: Detecting points

- We also count the points and give them an item number, which we will use to store the values in the matrix.

- Our last step is to generate a 3X21 matrix, we will generate three text files, generalinformation.txt, skeletonimage.txt, skeletonworld.txt. These three are the final output which stores our extracted features from the entire image.

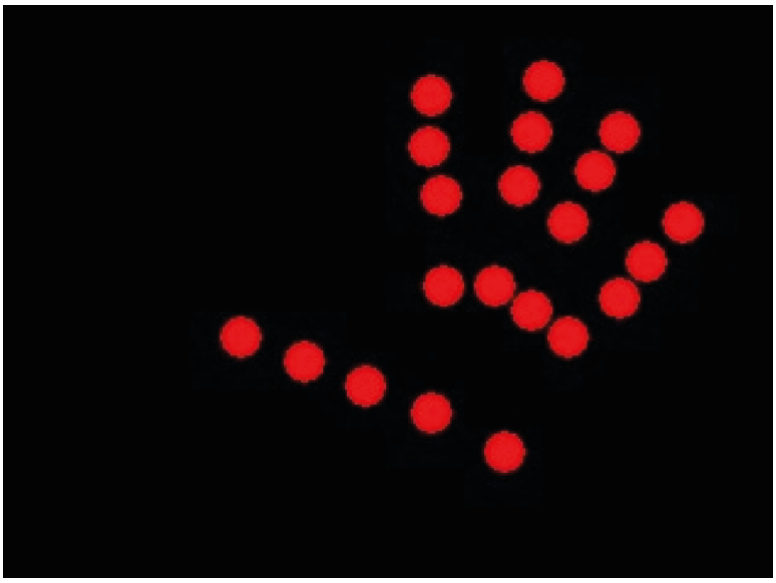

Figure 13: Subtracting background

4 Result Analysis

To get the coordinate of the hand points based of depth, intel realsense D435 was used, which was released in 2018. In this device, an image is captured on 3 axes, x, y and z. Here Z coordinate is basically the distance of the point from the sensor. We set the camera on the eye level of the participant and took each photos with depth information. Which is the raw data for the basic analysis.

4.1 Device and Environment Description

In this device the distance is measured by using the predefined distance between the sensor and light source. And finding the angle of the point of interest. Then by using mathematical formula, the distance is measured and the intensity change can be mapped based on the distance of object from the sensor. Python version 3.6 was

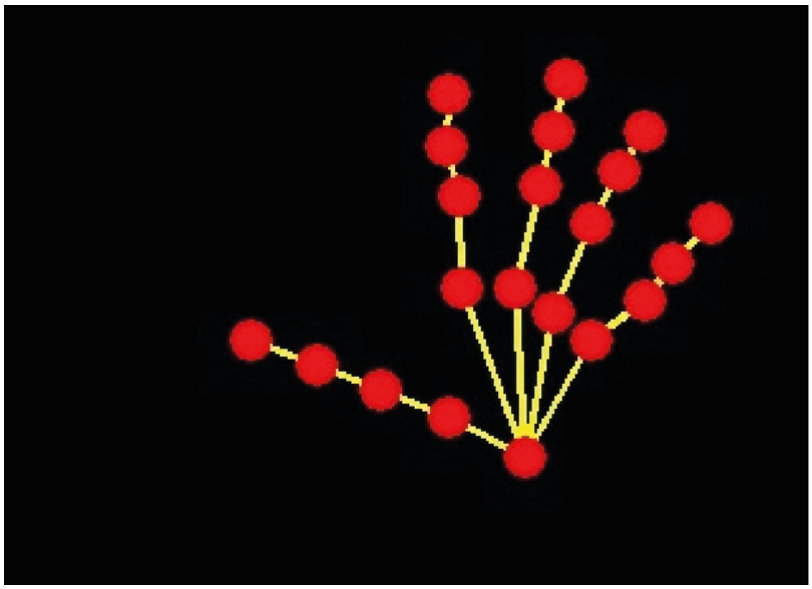

Figure 14: Skeleton of the hand

used with specific library like 'pyrealsense', 'PIL' etc. To capture the raw images, intel realsense SDK version 2.25 was used. The device was used to capture the images with raw information.

- Intel real Sense D435
- Processor: Intel Core i5-5820K @ 2.4 Ghz
- Chipset : Intel X99 Express Chipset
- Ram: 8 GB @ 2400 Mhz
- Platform: *python* 2016 64 bit

4.2 Experimental Result

Initially captured photos: 100 Plan to take photos: 15000 Information extracted from from images: Coordinates of the hand points. There are twenty one points detected from one hand. As two hands are used for a gesture, total 44 coordinates will be stored in a text

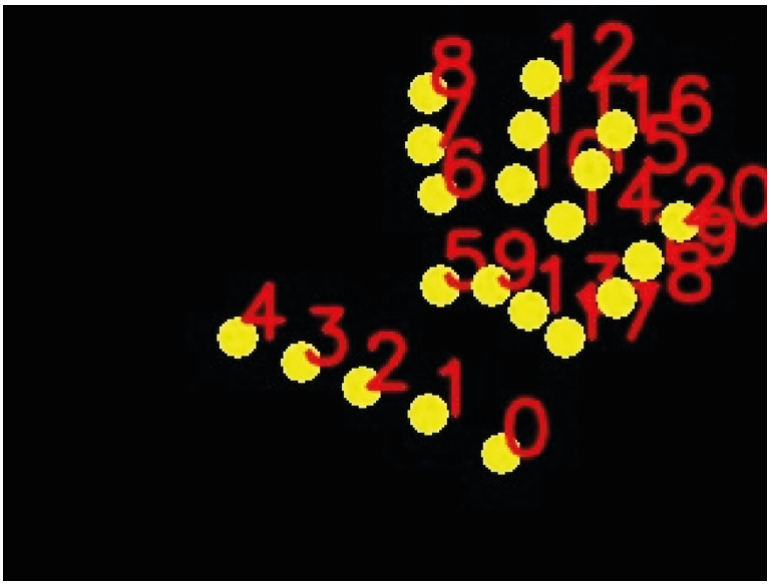

Figure 15: Numbering hands

file for each photo. The text file will contain each line containing coordinates of one gesture. There are separated files for each sign. We have separated the train data and test data for further training.

There were some challenges using this newly formed sensor. As the sensor is one of the latest, the documentation is not well defined. And there were not much examples regarding extracting the depth information. The examples were only based on projecting the images based on depth. Not extracting the depth values. That is why the model was used for extraction of the depth data.

5 Conclusion and Future Work

After all, our primary goal is to create a suitable data-set with reliable features, so that it can pave the path of future research based on Bengali sign language detection. Using depth image, we will create the data-set and then apply appropriate methods like deep learning or random forest to detect the alphabets. Our target is to target data from Bangladesh National Deaf and Dumb Association, where the professional interpreter will help us to collect the training data. Also, we will get data from people of different ages to find errors and solve them. The collection can be divided into three categories, professional interpreter, moderate interpreter and beginner. All these categories will help us to create the desired data set.

Currently we are working with vowels. Our plan is to collect data for the whole dataset. We are planning to capture around 2000 images for the alphabets and numbers. Then we are planning to go to dynamic gestures to detect dynamic sign language. Because dynamic gestures are used for regular conversation. Currently the smartphones contain depth camera sensors. If we can build a recognition system, that can detect the sign language gestures, then the boundary of communication with the mute and deaf will be eradicated.

We are planning to remodel the entire deeplearning model from scracth as soon as the intel realsense releases brand new library which they promised to release by the end of 2019. We believe with a brand

new model and more flexibility from our end would provide a more genuine scenario. We would be able to open new possibilities to utlise this assets for future work by providing them to the masses. And this will bring new possibilities to go one step further into developing flexible bangla sign language gesture recognition in future.

References

[1] Brashear, Helene and Henderson, Valerie and Park, Kwang-Hyun and Hamilton, Harley and Lee, Seungyon and Starner, Thad*American Sign Language Recognition* in Game Development for Deaf Children , 2006

[2] M. M. Sole and M. S. Tsoeu *Sign language recognition* using the Extreme Learning Machine , 2011

[3] Koller, O, Zargaran, O, Ney, H, Bowden, R *Deep Sign: Hybrid CNN-HMM* for Continuous Sign Language Recognition, 2016

[4] Ali Karami,Bahman Zanj,Azadeh Kiani Sarkaleh *Persian sign language (PSL) recognition* using wavelet transform and neural networks, 2011

[5] Tomas Simon, Hanbyul Joo, Iain Matthews, Yaser Sheikh *Hand keypoint detection in single images* using multiview bootstrapping, 2017

[6] S. K. Kang and M. Y. Nam and P. K. Rhee *Color Based Hand and Finger Detection* Technology for User Interaction, 2008 key7 Yue Wang, Dinggang Shen, Eam Khwang Teoh *Lane detection* using spline model, 2000

[7] J. Suarez and R. R. Murphy *Hand gesture recognition* with depth images: A review, 2012

[8] Brashear, Helene and Henderson, Valerie and Park, Kwang-Hyun and Hamilton, Harley and Lee, Seungyon and Starner, Thad*American Sign Language Recognition* in Game Development for Deaf Children , 2006

[9] M. M. Sole and M. S. Tsoeu *Sign language recognition* using the Extreme Learning Machine , 2011

[10] Koller, O, Zargaran, O, Ney, H, Bowden, R *Deep Sign: Hybrid CNN-HMM* for Continuous Sign Language Recognition, 2016

[11] Ali Karami,Bahman Zanj,Azadeh Kiani Sarkaleh *Persian sign language (PSL) recognition* using wavelet transform and neural networks, 2011

[12] Tomas Simon, Hanbyul Joo, Iain Matthews, Yaser Sheikh *Hand keypoint detection in single images* using multiview bootstrapping, 2017

[13] S. K. Kang and M. Y. Nam and P. K. Rhee *Color Based Hand and Finger Detection* Technology for User Interaction, 2008 key15 Yue Wang, Dinggang Shen, Eam Khwang Teoh *Lane detection* using spline model, 2000

[14] J. Suarez and R. R. Murphy *Hand gesture recognition* with depth images: A review, 2012

[15] Fabio Dominio,Mauro Donadeo,Pietro Zanuttigh *Combining multiple depth-based descriptors* for hand gesture recognition, 2014

[16] M. M. Sole and M. S. Tsoeu *Sign language recognition* using the Extreme Learning Machine, 2011

[17] Z. Ren and J. Meng and J. Yuan *Depth camera based hand gesture recognition and its applications* in Human-Computer-Interaction, 2011

[18] C. Wang and Z. Liu and S. Chan *Superpixel-Based Hand Gesture Recognition* With Kinect Depth Camera, 2015