Real Time Hand Gesture Recognition in Depth Image using CNN

Dardina Tasmere
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology

Boshir Ahmed
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology

Sanchita Rani Das
Department of Computer Science & Engineering
Prime University

ABSTRACT
Hand gestures can play a notable role in computer vision, and hand gesture-based methods can stand out in providing a native way of interaction. Deafness is a degree of loss such that a person is unable to understand speech, spoken language. Sign language declined the gap in spoken language. The hand gesture is analyzed identically to sign language presenting the naturalness of intercommunication for deaf people. Real-time hand gesture recognition has been proposed in our research. Our proposed CNN model architecture will remediate the communication barrier of deaf people. The proposed model has achieved an accuracy of 94.61% to recognize 11 several different gestures using depth images.

General Terms
Hearing-impaired people, Computer Vision, Hand gesture recognition

Keywords
Depth Image, Deep Convolution Neural Network, Real Time Recognition

1. INTRODUCTION
Disability is an essential subject concerning human rights. Disability can be categorized into different types: hearing disability, visual disability, physical disability, speech disability and mental disability. Hearing stands as one of the essential senses of individuals. A hearing loss person is incapable of hearing that is hearable to others. Hearing loss leads to complexity well as it may be mild, moderate, severe, or profound. According to the world health organization, more than 5% of the world's people – approximately 466 million people – have hearing impairments (432 million adults and 34 million children) [1]. According to an assessment in 2050, one in every ten people will have hearing loss problems. Hearing impairment drives people to be excluded from fundamental communication, thereby adding social isolation, frustration. Hearing impaired and deaf people apply sign language as a helpful meaning of communicating.

Sign language is used as a medium to express thinking includes a different body or arms postures, a combination of several hand position and shape, facial expressions. Sign language declined the gap in spoken language. Sign language is not universal. However, there exist distinct correlations sign languages vary from country to country and culture to culture. There are different sign languages like American sign language, Japanese sign language, Indian sign language, Arabic sign language, and so for respective countries also.

Without communication, people cannot imagine a single instant. Different technical aid is supporting the communication process. But physically disabled persons do not get the advancement support of technical improvements. Normal people do not take it easy to communicate with sign language users because of insufficient attention to communicate with deaf and dumb people. Hearing loss has a meaningful impression on a child's cultural improvement and educational achievements. Deaf learning has focused in many developed countries, but development in this field is still needed in developing countries as Bangladesh. In Bangladesh, deaf and dumb people are an unreached people group because of paying a lack of importance in society. As a result, sign language users can not enjoy sufficient access to information and services like education, health services, and employment opportunities.

Using excellent technological support, large amounts of the research project are applied to hand gesture recognition systems. This framework is designed to remediate the communication hindrance between deaf people and normal people.

2. RELATED WORKS
The hand gesture is one medium of signing. In computer vision and deep learning, gesture recognition has become one of the most interesting topics. Firstly the glove-based system was developed to recognize hand gestures, but the interface restricts the user comfort because of wearing heavy devices and the burden of cable connections [2]. Though providing low-cost hardware, portability facility, and charger transfer touch sensor, the proposed device had due to lack of naturalness [3]. For reaching out to the problem of impaired people, a considerable number of research projects are dedicated to translating ASL (American Sign Language). 3D depth-sensing camera- Kinect was used for image acquisition together with applying depth thresholding, contour, convex hull, and convexity defects were applied to the binary image for sign language classification [5]. Shukla et al. used the Naïve Bayes classification algorithm to classify five fingers only. In [6] the color skin information is used to segment the hands for the generic hand-based remote control system. Commonly hand gesture recognition system follows three phases: detection, tracking, and recognition. Survey [4] provides a perception of existing methods of gesture recognition systems for human-computer interaction by categorizing different key parameters.

In a couple of years, deep learning approaches have shown a flawless achievement [7] to suppress traditional machine learning techniques in computer vision. Peijun Bao et al. [7] proposed deep CNN to directly recognize the gesture from the whole image without using any region proposal algorithm or sliding window mechanism. On the other hand, recognition of gestures with both written words and audible speech was proposed [8] by image matching technique and use of build-in

"Find" function. Preceding analysis on different sign languages such as American, Indian [9], and Arabic, as well as Italian, has fascinated researchers to work with Bangla sign language. Rahman et al. proposed Haar-like feature-based cascade; Hue and Saturation implemented for hand sign extraction, and K-Nearest Neighbors (KNN) algorithm for 36 Bangla characters recognition yielded an average accuracy of 96.46% [10]. A real time sign language has been proposed for the ILSVRC2012 dataset for American Sign Language using GoogleNet architecture [11]. As human activity and gesture recognition has become one of the growing domains of ambient intelligence, the authors proposed convolutional neural networks (CNN) and the recurrent neural networks (RNN), for automated hand gesture recognition using both depth and skeleton data. Their method achieved overall accuracy of 85.46% for 14 dynamic gestures [12]. As main focus on depth images , so in [15] using Senz3d dataset depth images, real time hand gesture recognition was proposed by Memo et al. Hand was segmented by hand contour and multi-class Support Vector Machine was used to classify 11 gestures that gained accuracy 90% [15] .

This research objective is to facilitate the learning and communication process of deaf and hearing-impaired people.

3. DATASET DESCRIPTION

Solving challenging problems of a research dataset has a significant role. A dataset is particularly necessary for excellent research to gain more immeasurable certainty. Dataset importance is very apparent in deep learning. It is the most crucial aspect that makes algorithm training possible [20]. A large dataset plays a notable role in improving the classification rate in deep learning. To work with depth images, Senz3d dataset of hand gestures has been used [13][14] [15].

Fig 1: Sample Images from Dataset [13][14]

The dataset includes various different static gestures obtained by the Creative Senz3D camera [13] [14]. The dataset consists of 1320 sample images. The dataset is composed of four different people images. Each person made 11 different gestures, repeatedly for 30 times. In the dataset, for each sample, color, depth, and confidence frames are available. A snapshot of few images from the dataset for hand gesture recognition is given in Fig 1.

4. PROPOSED METHODOLOGY

In our research, the hand gesture recognition system has been proposed by hand segmentation and preprocessing operation (resizing all images) followed by CNN model architecture. However, image classification models have become currently prominent for computer vision field. After acquiring the depth image, segmented the hand data. Following some preprocessing activities, trained CNN model for feature extraction and gesture recognition. In Fig 2 the block diagram of proposed methodology has been shown.

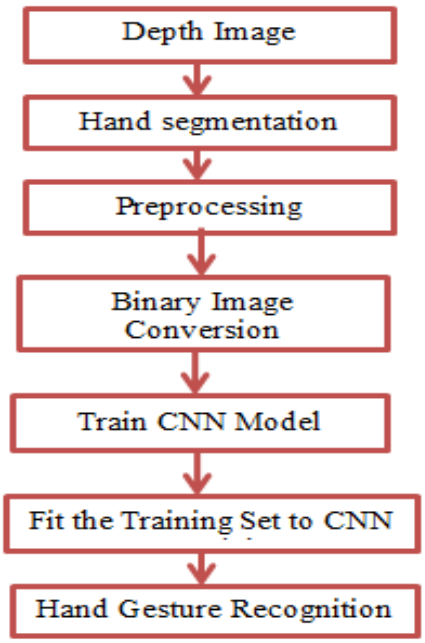

Fig 2: Proposed Methodology of the Framework

4.1 Hand Segmentation

To facilitate our work, depth images have taken from the dataset. In hand gesture recognition, extracting the hand region is the foremost task. So our first step towards gesture recognition is to segment the hand region from the depth map. The segmentation process implies that the hand is close to the camera and begins with thresholding of the depth values that eliminates the samples with a distance larger than a pre-defined threshold that is based on the selected application [23]. After acquiring the depth images, YC_bC_r color space has been implemented to segment the hand data. It clearly distinguishes the hand from the frame. The depth image has converted to YC_bC_r value. The YC_bC_r value of every pixel has compared with the standard values. A predefined threshold value for each parameter has given below:

Y, C_b, and C_r values range are as like:

$$0 \le Y \le 255 \text{ and } 135 \le C \le 180 \text{ and } 85 \le C_b \le 135.$$

4.2 Preprocessing

It is challenging to know how to properly prepare image data when training a convolutional neural network [19]. Images in the training dataset had differing sizes; therefore images had to be resized before being used as input to the model [19]. This involves both resizing and cropping techniques during both the training and evaluation of the model [19]. After getting the segmented hand images, the images have been resized to 256×256.

4.3 Binary Image Conversion

The binary image conversion approach helps the classification steps. It is quite easier to deal with binary image than multiple color channels. Thresholding transforms images into binary images. Pixels below the threshold value (60) have converted to 0 and pixels above the threshold value have converted to 1.

5. PROPOSED CNN MODEL

Artificial Intelligence is a witness to an impressive expansion

in bridging the gap between human and machine [20]. The changes in Computer Vision as well as deep Learning has become completed with time, fundamentally over one particular algorithm — a Convolutional Neural Network [20]. It has decisively proved over time that neural networks outperform other algorithms in accuracy and speed. Image classification has become a topic of current interest to researchers with the advancement of neural networks. Convolution neural network outperforms in this field than traditional machine learning methods. Fig. 3 represents a complete work flow of CNN to process hand gestures.

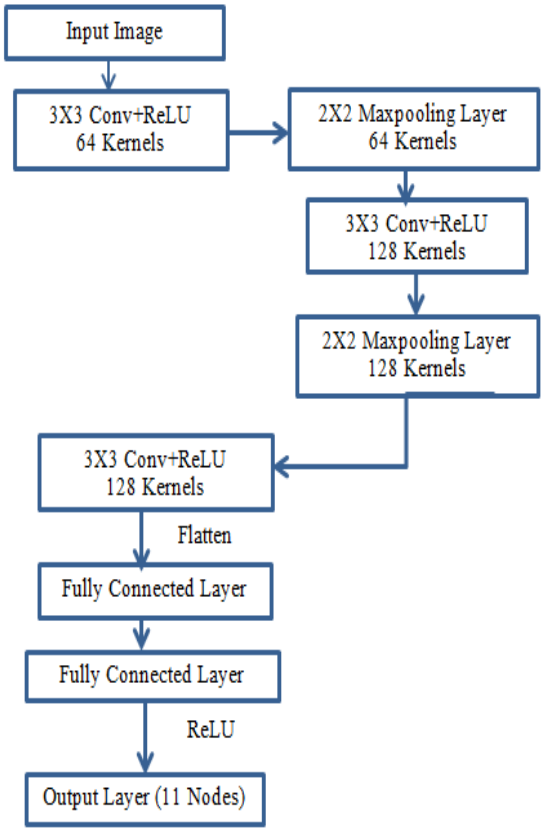

Fig 3: Architecture of proposed CNN model

Each input image will pass through a list of convolution layers, Pooling, moreover fully connected layers, and employ ReLU function to classify hand gestures with probabilistic values. In our proposed CNN architecture, there consists of three convolution layers, two maxpooling layers and two fully connected layers. The output layer consists of 11 nodes. These 11 nodes have proposed for recognizing 11 gestures from the dataset. From the dataset 816 images have been taken for training and rest of the images for validation.

5.1 Convolution layer

The normalized images have fed to the convolution layer. The input images resized to 256×256. The input image has proceeded by four convolution layer with a tractable feature map. A kernel can be regarded as an array of numbers often termed as weights. In the convolution layers, the kernel size has selected 3×3. A mathematical operation has conducted upon the image by the kernel. Convolution of an image with various filters performs distinctly as edge detection, blur, and sharpness. In our proposed methodology, three convolution layers has used. First convolution layer has used 64 filters but in second and 3rd convolution layer 128 filters have chosen. Filters in convolutional layers detect features that help classification.

5.2 Pooling Layer

The pooling layer, similar to the first layer reduces the number of parameters, spatial size of features. Pooling lessens the dimensionality of feature maps but holds significant features. Implementation pooling on each feature map separately, the same number of pooled feature maps is created. In our proposed architecture, two pooling layers have used. Pooling includes choosing a pooling formula. Among different pooling types, max-pooling has chosen. Max pooling has been found to work better in usage than other pooling operation for computer vision studies like image classification. The results are pooled feature maps that highlight the most present feature in the patch, not like the case of average pooling. Maximum pooling is a pooling operation that calculates the maximum value in each section of each feature map. Kernel size is always smaller than the feature map. The result of the pooling layer is a summarized version of the features identified from the input frame. Small changes in the location of the feature in the input detected by the convolutional layer will result in a pooled feature map. The pooling layer reduces the computational power needed to process images by dimensionality reduction. Moreover, it is beneficial for extracting powerful features that are rotational and positional invariant to maintain the training process more effective.

5.3 Fully Connected Layer

The last layer of the convolution neural network is the fully connected layer. These layers act identically like a traditional deep neural network. Every neuron in one layer connects to every neuron in another layer through the fully connected layer. Convolution, as well as pooling layer, does feature extraction, whereas fully connected layers do classification based on the features by the previous layer. The FC layer holds composite information from all the convolution and pooling layers. The flattened matrix goes through an FC layer to classify the images. The feature vector representation is done by a fully connected layer (FC). This feature vector carries necessary information about the input. When the network gets trained, this feature vector is then further used for classification. Based on learning non-linear functions, this layer classifies images based on features. This layer contains a ReLU activation function, which provides a probability for each of the classification labels.

5.4 ReLU Activation Function

In the neural network, the activation function is responsible for transforming the summation of weighted input [21]. The rectified linear activation function **(**ReLU) gives direct output otherwise zero. This linear function provides output the input directly if it is positive, otherwise, it will output zero [21]. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance [21].

6. RESULT ANALYSIS

This section of the study broadly analyzes experimental results of the proposed framework of recognizing 11 gestures for real time video. The model has trained on the Senz3d dataset. Our target is to keep errors low as possible using the proposed methodology. Accuracy curve measures how the model is effective in terms of accuracy. The accuracy curve of the proposed model for hand gesture recognition is given below in Fig 4. The accuracy has converged near to 94.61 for 11 gestures.

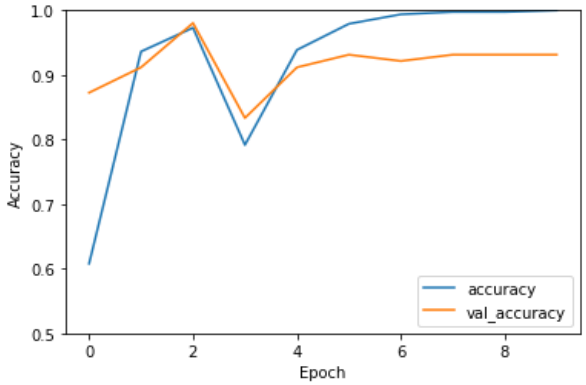

Fig 4: Accuracy curve for real time hand gesture recognition in depth image

To compile our proposed methodology, the batch size was 32 and seed value has been selected as 32. The process has completed compilation in 10 epoches.

Accuracy finds how correct the proposed methodology is on predicted conditions. Accuracy is helpful to resolve if the charges of false positives are unusual. Recall defines the actual positives of the proposed methodology by classifying gestures. The performance of the designed architecture for hand gesture recognition in terms of accuracy has given below in Table 1.

Table 1: Performance of the proposed architecture hand gesture recognition

Class	Accuracy
0	0.9375
1	0.9374
2	0.9642
3	0.9774
4	0.9841
5	0.9542
6	0.9458
7	0.9462
8	0.9647
9	0.9420
10	0.9841

In our gesture recognition system, webcam has used to validate our proposed gesture recognition system. Due to lighting effect, output value has fluctuating. Various input has been given such as one, two, four gestures before webcam, the gesture recognition model has given output according to input gesture from the webcam. Snapshots of real time hand gesture recognition system using proposed methodology have sketched out in Fig. 5 for several gestures.

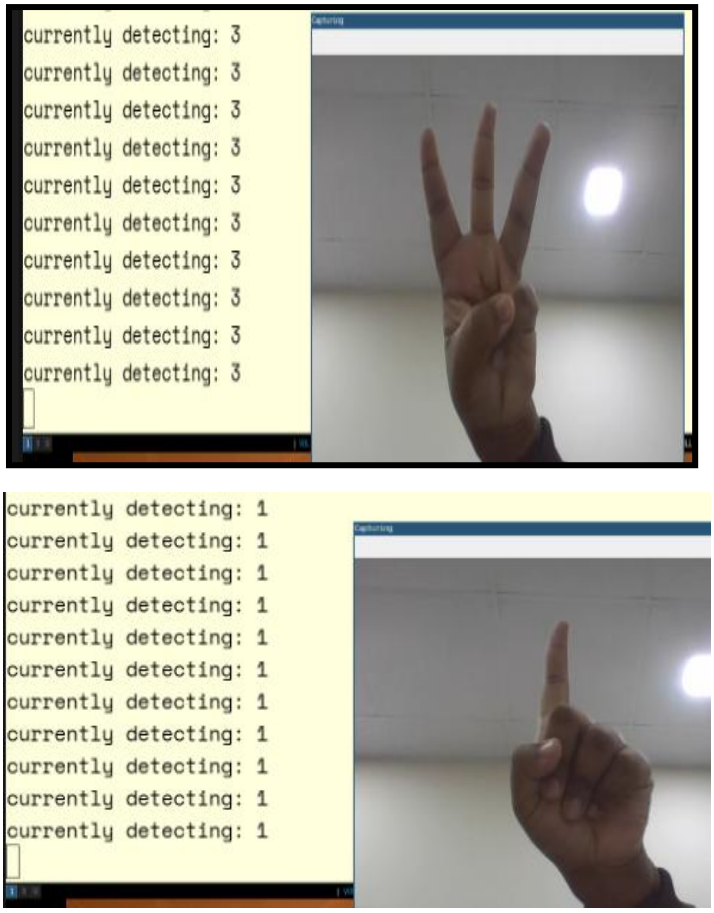

Fig 5: Output of real time hand gesture recognition using proposed methodology

6. CONCLUSION

In this work, a real-time hand gesture recognition system has been proposed to overcome the communication challenges of deaf people. Real-time gesture recognition from depth images may introduce a new way to this research area. For gesture recognition, YC_bC_r color space has been used for hand segmentation following by the proposed CNN model. The proposed CNN model includes three convolution layers, two max-pooling layers, and two fully connected layers. This proposed methodology has given accuracy of 94.61% for 11 gestures from depth images. To obtain greater efficiency, the Senz3d dataset with 1320 sample images has been used. In the future, a background subtraction model can be used for hand segmentation from depth images to implement a realistic real-time hand gesture recognition system.

7. REFERENCES

[1] W. H. Organization et al. 2020 Deafness and hearing loss fact sheet n 300. updated March,2020" 2020.

[2] Pavlovic, V.I., Sharma, R. and Huang, T.S., 1997. Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on pattern analysis and machine intelligence, 19(7), pp.677-695.

[3] Abhishek, K.S., Qubeley, L.C.F. and Ho, D., 2016, August. Glove-based hand gesture recognition sign language translator using capacitive touch sensor. In 2016 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC) (pp. 334-337). IEEE.

[4] Rautaray, S.S. and Agrawal, A., 2015. Vision based hand gesture recognition for human computer interaction: a survey. Artificial intelligence review, 43(1), pp.1-54.

[5] Shukla, J. and Dwivedi, A., 2014, April. A method for hand gesture recognition. In 2014 Fourth International Conference on Communication Systems and Network Technologies (pp. 919-923). IEEE.

[6] Erden, F. and Cetin, A.E., 2014. Hand gesture based remote control system using infrared sensors and a camera. IEEE Transactions on Consumer Electronics, 60(4), pp.675-680.

[7] Bao, P., Maqueda, A.I., del-Blanco, C.R. and García, N., 2017. Tiny hand gesture recognition without localization via a deep convolutional network. IEEE Transactions on Consumer Electronics, 63(3), pp.251-257.

[8] Akoum, A. and Al Mawla, N., 2015. Hand gesture recognition approach for asl language using hand extraction algorithm. Journal of Software Engineering and Applications, 8(08), p.419.

[9] Akoum, A. and Al Mawla, N., 2015. Hand gesture recognition approach for asl language using hand extraction algorithm. Journal of Software Engineering and Applications, 8(08), p.419.

[10] Rahaman, M.A., Jasim, M., Ali, M.H. and Hasanuzzaman, M., 2014, December. Real-time computer vision-based Bengali sign language recognition. In 2014 17th International Conference on Computer and Information Technology (ICCIT) (pp. 192-197). IEEE.

[11] Garcia, B. and Viesca, S.A., 2016. Real-time American sign language recognition with convolutional neural networks. Convolutional Neural Networks for Visual Recognition, 2, pp.225-232.

[12] Lai, K. and Yanushkevich, S.N., 2018, August. CNN+ RNN depth and skeleton based dynamic hand gesture recognition. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 3451-3456). IEEE.

[13] G. Marin, F. Dominio, P. Zanuttigh, "Hand gesture recognition with Leap Motion and Kinect devices", IEEE International Conference on Image Processing (ICIP), Paris, France, 2014

[14] G. Marin, F. Dominio, P. Zanuttigh, "Hand Gesture Recognition with Jointly Calibrated Leap Motion and Depth Sensor", Multimedia Tools and Applications, 2015.

[15] Minto, L. and Zanuttigh, P., 2015. Exploiting silhouette descriptors and synthetic data for hand gesture recognition.

[16] Santosh, K.C. and Hegadi, R.S. eds., 2019. Recent Trends in Image Processing and Pattern Recognition: Second International Conference, RTIP2R 2018, Solapur, India, December 21–22, 2018, Revised Selected Papers, Part I (Vol. 1035). Springer.

[17] Hossain, S., Sarma, D., Mittra, T., Alam, M.N., Saha, I. and Johora, F.T., 2020, July. Bengali Hand Sign Gestures Recognition using Convolutional Neural Network. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 636-641). IEEE.

[18] Islalm, M.S., Rahman, M.M., Rahman, M.H., Arifuzzaman, M., Sassi, R. and Aktaruzzaman, M., 2019, September. Recognition bangla sign language using convolutional neural network. In 2019 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) (pp. 1-6). IEEE.

[19] Brownlee,J.,2019. Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python. Machine Learning Mastery

[20] Saha, S., 2018. A comprehensive guide to convolutional neural networks—the ELI5 way. Towards Data Science, 15.

[21] Brownlee, J., 2019. A gentle introduction to the rectified linear unit (relu). Machine Learning Mastery. https://machinelearningmastery.com/rectified-linear-activation-function-fordeep-learning-neural-networks.