American Sign Language Detection Using Convolutional Neural Network

Syeda Rabita Alam
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
syeda.alam10@northsouth.edu

Tamalika Bakshi
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
tamalika.bakshi@northsouth.edu

Sariah Tul Qarim
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
sariah.qarim@northsouth.edu

***Abstract*— A word delivery system that will serve people as a learning or detecting tool in which to help with hand movement. Usually, this project will help disabled humans who cannot hear or talk. However, we can define sign language by communication utilizing hand gestures and movements, presenting and facial expressions, rather than the spoken word. Although sign language is present in recent times, it is challenging for ordinary people to understand or communicate with sign language speakers. But nowadays, with the advantage of artificial intelligence and deep learning, there have been ensuring progress models in image recognition using deep learning and computer vision-based algorithms. This work aims to create a deep learning-based application that offers sign language translation to words or letters to build a communication system between signers and non-signers. The proposed model holds an ImageNet dataset and extracts them to a hyperspectral image's spatial features. After that, we can use a CNN (Convolutional Neural Network) model for identifying spatial features.**

Keywords*— *Non-verbal, Fundus image, CNN, deep learning, pre-processing, Dataset, hyperspectral image

I. INTRODUCTION

Sign language could be a variety of communication utilized by individuals with impaired hearing and speech. Individuals use linguistic communication gestures as a non-verbal communication to specific their thoughts and emotions. Not only the disabled are use sign language to express themselves, but ordinary people use this language to communicate with them. However, non-signers cope with it very difficult to know, thence trained linguistic communication interpreters are required throughout medical and legal appointments and instructional and coaching sessions. It is further complicated to communicate between the Deaf community and the hearing majority. The alternative solution is handwritten to communicate. Handwritten is more troublesome for Deaf-Community as they are generally less skilled in writing a spoken language. [1]. Furthermore, this type of communication is impersonal and slow in face-to-face conversations in emergency situations. Another challenge of learning sign-linguistic communication is that the expected speed of communication in a social setting will usually be overwhelming, significantly if we are in social groups. It takes contacts to a completely different level and demands that you just master eye gazing to navigate the give-and-take of communal interactions. The purpose of this work is to contribute to the field of automatic sign language recognition. We focus on the recognition of the signs or gestures. The objective of the aim of our project is to develop a machine architect Monitoring System. We are developing machine learning and computer vision applications. For completing this task there would be some challenges like Image Data pre- processing, identifying the best hyper parameter for the CNN Model etc. An RNN (Recurrent Neural Network) model can be used for extract temporal features from the hyperspectral image via two methods: Using the outputs from the SoftMax and the Pool layer of the CNN respectively. Afterwards, we also use the loss estimation algorithm to identify the accuracy. Our proposed dataset has different gestures performed one sign language multiple times giving us variation in context and color variation in different angel as well as.

Convolution neural network is one of the most common techniques to process shapes inside the image. With the help of this we can detect and classify an image or a specific portion of the image. This is a very useful method and easy to implement. Our input image is compressed using feature extraction and soon it is converted into a linear 1D array [2]. Our research is a very basic experiment of sign language to determine the expression. We are not going to use a large scale of data and instead of that we are taking a small number of images as instructed. A scratch model is developed. AS a result, the accuracy may fluctuate in different cases but in raw preprocessed images it will be able to show the proper classification of the image.

Deep learning-based simulation approaches with so many processing layers to acquire multiple levels of abstraction for data representations. Our research is unsupervised learning because in this case our class mode is categorical instead of binary which means output of our research is not yes-no type rather is says the class of the image.[3]

As soon as the signed languages are accepted as human languages, a whole new universe of possibilities opens up. The nature of human language and language processing, the relationship between cognition and language, and the brain organization of language may all be studied using signed languages. The modality of sign languages is what makes them valuable. Sign languages rely on high-level vision and motion processing systems for perception, and they require the integration of motor systems including the hands and face for production.[4]

Durability and permanence Nonverbal greetings are those that do not require speech or voice, and can be understood by entire classes rather than individuals. The majority of non-vocal greetings are expressed through

physical gestures. According to dense linguist requirements, these have been mentioned throughout the delivery note. Sign language is used by disabled men and women to communicate with their talking but hard-of-hearing partners, although greetings with gestures are used by both regular and broken individuals. So, since we do speak, greetings with gestures are a common occurrence, as evidenced by research.[5]

II. LITERATURE REVIEW

Literature review of the problem shows that there have been several approaches to address the issue of gesture recognition in images or video using several different methods. In [6] the Authors showed two different approaches using LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Network). This proposed model takes video sequences and extracts temporal and spatial features from them. They used Inception, a CNN for recognizing spatial features, and an RNN (Recurrent Neural Network) to train on temporal features. The dataset used is the American Sign Language Dataset. They took the outputs of the SoftMax Layer and the Max Pooling Layer where the got more accuracy with SoftMax about 90-93% than Pool Layer about 55-58%. One of the problems the model faced is with facial features and skin tones. While testing with different skin tones, the model dropped accuracy if it hadn't been trained on a certain skin tone and was made to predict on it.

In [7] also published a paper on a skin-color modeling technique. The skin-color range is predetermined that will extract pixels (hand) from non–pixels (background). The images were fed into CNN for classification of images. Keras was used for training of images. Provided with proper lighting condition and a uniform background, the system acquired an average testing accuracy of 93.67%, of which 90.04% was attributed to ASL alphabet recognition, 93.44% for number recognition and 97.52% for static word recognition, thus surpassing that of other related studies. The approach is used for fast computation and is done in real time. The authors declare no conflict of interest.

In [8] a paper on American Sign Language Alphabet Recognition proposing a model to recognize American Sign Language alphabet from RGB images. Images for the training were resized and pre-processed before training the Deep Neural Network. The model was trained on a squeezenet architecture to make it capable of running on mobile devices with an accuracy of 83.29%. But they observed that this happens when similar looking alphabet like 'a' and 't' where the difference between them is thumb on the side for 'a' whereas 't' has thumb in between index and middle finger. When an image with different light conditions is given, or the fingers are not visible then it leads to a false prediction.

In another paper [9] exhibits the assessment of different pixel level highlights for the dual handed sign language dataset. The element extraction techniques are Histogram of Orientation Gradient (HOG), Histogram of Boundary Description (HBD) and the Histogram of Edge Frequency (HOEF). The exactness of HOG and HBD found up to 71.4% and 77.3% while the precision of HOEF, all things considered, informational collection is 97.3% and in perfect condition 98.1%.

[10] The authors tried to convert ISL (Indian Sign Language) to English Language sentence using three approaches. First using an LSTM based Sequence to Sequence model (Seq2Seq), second using an LSTM based Seq2Seq model utilizing attention, third using an Indian Sign Language Transformer. These models were evaluated on BLEU scores and the transformer model gave a perfect BLEU score of 1.0 on test data where the first two approaches gave score of 0.59 and 0.67 respectively.

The paper [11] on ISL recognition deals with robust modeling of static signs using CNN. The efficiency of the proposed system is evaluated on approximately 50 CNN models. The results are also evaluated on the basis of different optimizers- ADAM and RMSProp, and it has been observed that the proposed approach has achieved the highest training accuracy of 99.72% and 99.90% on colored and grayscale images, respectively.

However, all of these papers are mainly based on either alphabets or static words from real time. But our focus is to solve this problem in an easier way where the words will be detected from images using simpler CNN so that it becomes more efficient and user friendly. Our method is better than these methods, therefore our aim is to solve this problem for those who need it most and can use it very easily.

III. METHODOLOGY

The key objective of our work is to design and develop a model that can classify Traffic Signs. Figure 1 shows the schematic diagram of the proposed . For our proposed methodology, raw image will be regarded as input data. The whole dataset is split into training ,test and validation dataset. In the preprocessing phase, data are represented in such a way that features can be mapped from them. The features which are extracted using CNN model are transformed into features vector. Then output of the model determines particular sign.

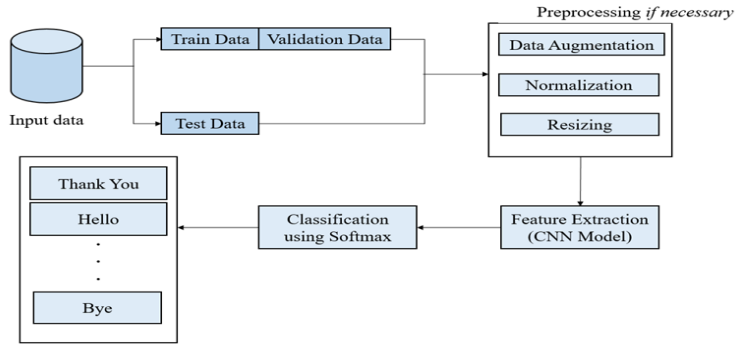


Figure 1: Proposed Methodology for Sign Language Classification

A. *Dataset Description*

Dataset will be described when the project is completed.

B. *Convolutional Neural Network (CNN)*

CNNs are deep FNNs which learn hierarchies of invariant features automatically. Features are distinctive attributes or aspects of something and invariant in the context of CNNs means that the same transformation is applied at different locations.

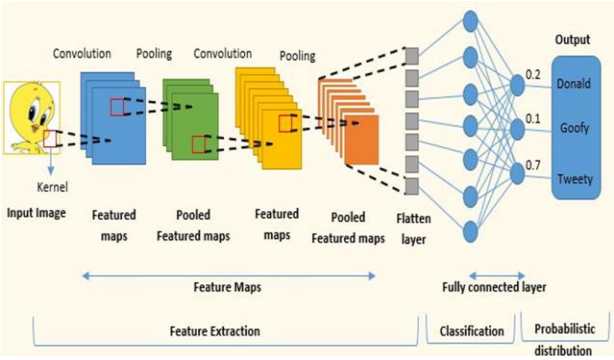


Figure 2: Proposed Methodology for Traffic Sign Classification

C. *Convolutional Layer*

A CNN consists of several convolutional layers which are used to get the features extracted from the input of the network. A key concept of CNNs is that the same transformationis applied at all locations.

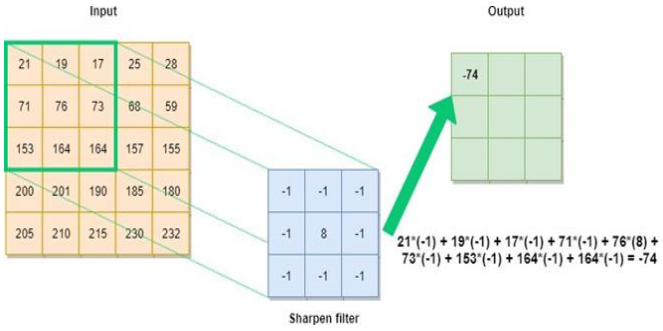


Figure 3: Convolutional Layer

D. *Feature Extraction*

The qualities related to the context of the scenes are characterized as picture spatial feature extraction. CNN has recently demonstrated its ability to extract complicated spatial characteristics from photos to a significant extent. The convolutional layer, which is named after the network, is a crucial component of the CNN. Convolutional layers may "convolution" pictures to discover local and translation invariant patterns. It refers to a linear mathematical procedure that conducts matrix multiplication between a filter of a specified dimension and the picture region over which the filter is hovering. A "feature map" is created by passing the output of a convolutional layer to an activation function, which introduces nonlinearity. Another integral element of CNN is the pooling layer that has come up with a down sampling strategy.

E. *Pooling Layer*

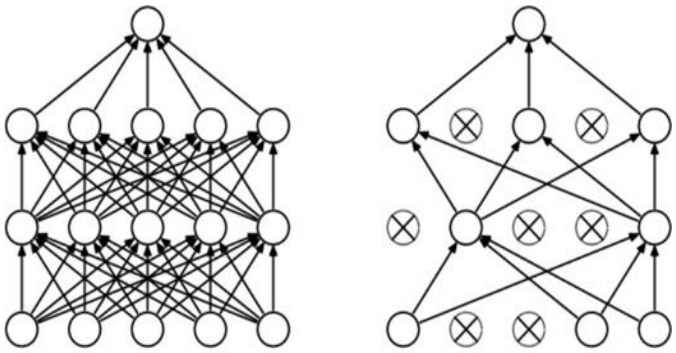
In the pooling layer in a CNN, a volume is down sampled spatially. The down sampling is done by itself in each portion of input volume. The idea of pooling in CNNs is basically to decrease parameters in the connection of neurons and to make the feature detection immune to scale and orientation changes. In other words, pooling generalizes the lower level information and enables moving from high resolution data to lower resolution information.

F. *Fully Connected Layer*

The fully connected layers make classifications of the objects from the convolution pooling layers' output. In short, one can say that it is a standard neural network classifier attached onto the end of a high level feature extractor. The final pooling layer's output is passed through many channels of X × Y matrices and to be able to connect the pooling layer with the fully connected layer the output needs to be flattened to a single matrix of size 1*N.

G. *Droput*

A recently introduced technique, dropout, addresses the issue with not having enough training data and it prevents over fitting. Dropout means that some units (hidden and visible) in the neural network are dropped out temporarily, that is, are removed from the network. The dropped out units do not take part in the forward pass and back propagation. This makes the neural network sample a different architecture every time an input is presented. Dropout reduces complex co-adaptations of the units in the neural network, since a unit cannot rely fully on the presence of particular other units since they might be dropped out. This makes the network more robust.



(a) Standard Neural Net (b) After applying dropout.

H. *Classification Using SoftMax*

For classification, the SoftMax activation function is used in the output layer that pro- duces the probability

distribution of the five classes based on the extracted features of the previous step. This function outputs probabilistic values ranging between 0 to 1, all summing up to 1. SoftMax activation function generates probability distribution by using the following equation:

$$S(i)_k = \Sigma [e\ ik /n=1^t e^{in}]$$

In this equation, → i denotes the input vector, k is the index of the current element in the input vector, all the i values refer to the elements of the input vector, and t representsthe total number of classes.

I. Scratch Model

Constructing a deep neural network from the bottom up is essential for gaining a better understanding of deep learning methods and insight into the dataset's feature space. A scratched model is developed in this research with including CNN. The transformed images are passed to the CNN architecture, that is deployed as a feature extractor, once the preprocessing processes are completed. Convolution layers with a varying set of parameters of 5X5 and 3 X3 dimensionality, maxpool of 2X 2 pool size, and flatten layers are all part of the CNN architecture. ReLU is implemented as an activation function that outputs to the [0,] range for faster convergence while training. Afterwards, a flatten layer is appended to create a single long feature vector for each selected image. A dense layer of X and Y neurons is added next with a dropout rate of 0.5, followed by the output layer, where X and Y will be defined later and the Softmax activation function is used to classify 5 sign language classes. Overall steps of the Scratch model architecture is shown below in figure 4.

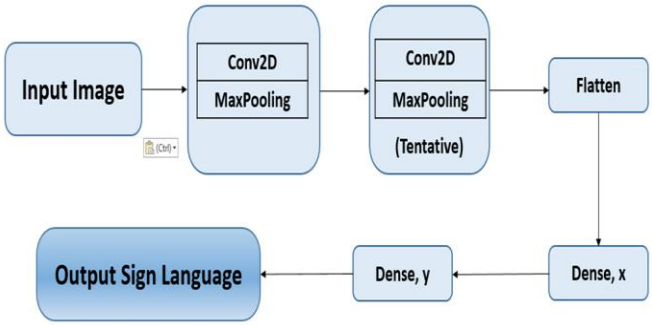


Figure 4: Scratch Model Architecture

IV. Future Work

Our dataset will be extremely small and more images should be added to the train the model better. Besides, we can add transfer learning to improve performance. A web-based app can be developed so that people from all over the world can translate sign language to human language.

V. Conclusion

Sign language is one of the primal techniques to communicate with each other. Nowadays sign language has become more popular and deaf people need to communicate with others. As most of the time normal people is not familiar with sign language, so they can't communicate if they know sign language. Our research will be helpful in this condition to communicate with each other.

References

[1] Van Herreweghe, M.: Prelinguaal dove jongeren en nederlands: een syntactisch onderzoek. *"Universiteit Gent"*,volume 51. Faculteit Letteren en Wijsbegeerte ,1996

[2] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In 2017 international conference on engineering and technology (ICET), pages 1–6. Ieee, 2017

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, 2015

[4] Karen Emmorey. Language, cognition, and the brain: Insights from sign language research. Psychology Press, 2001

[5] Najlaa Hayyawi Abbar and Hasanain Hassan Shaheed.The difference between american sign language and body language in greetings. Multi-cultural Education, 7(5), 2021

[6] Kshitij Bantupalli, Ying XieJ, "American Sign Language Recognition using Deep Learning and Computer Vision", IEEE, 2018

[7] Lean Karlo S. Tolentino, Ronnie O. Serfa Juan, August C. Thio-ac, Maria Abigail B. Pamahoy, Joni Rose R. Forteza, and Xavier Jet O. Garcia, "Static Sign Language Recognition Using Deep Learning", International Journal of Machine Learning and Computing, Vol. 9, No. 6, December 2019

[8] Nikhil Kasukurthi, Brij Rokad, Shiv Bidani, Aju Dennisan, "American Sign Language Alphabet Recognition using Deep Learning", IEEE, May 2019

[9] Lilha, H., & Shivmurthy, D. (2011, December). Analysis of pixel level features in recognition of real life dual-handed sign language data set. In Recent Trends in Information Systems (ReTIS), 2011 International Conference on (pp. 246-251). IEEE

[10] Pratik Likhar; Rathna G N, "Indian Sign Language Translation using Deep Learning", IEEE, 2021

[11] Ankita Wadhawan, Parteek Kumar, "Deep learning-based sign language recognition system for static signs", Springer, January 2020