CNN based Sign Language Recognition System with Multi-format Output

2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) | 979-8-3503-3086-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAC3N60023.2023.10541724

Harshit Pandey
Department of Electrical Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
harshit110801@gmail.com

Amaan Ahmed
Department of Electrical Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
md.aman26@gmail.com

Tushar Kumar
Department of Electrical Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
tushar007665@gmail.com

Vaibhav Kumar Singh
Department of Electrical Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
vksvaibhav420@gmail.com

Lipika Dutta
Department of Electrical Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
lipika.datta@galgotiacollege.edu

Pinki Yadav
Department of Electrical Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
pinki.yadav@galgotiacollege.edu

***Abstract*—Despite being one of the oldest and most natural forms of communication, sign language is challenging to comprehend since very few people are conversant in it. In this article a real time method for sign recognition using neural networks for finger spelling based on American sign language is presented. The purpose is to recognize hand gestures of human task activities from a camera image. The position of hand and orientation are applied to obtain the training and testing data for the CNN after applying several imageprocessing techniques. The hand is first put through a filter, and once that has been done, it is put through a classifier, which determines what class the hand movements belong to. Then the calibrated images are used to train CNN. Further various performance parameters such as accuracy, precision, recall and f1 score are calculated to analyze the proposed model. Using the DenseNet-169 CNN model the proposed system showed excellent performance with an accuracy of 99%. Moreover, the output is obtained in both text as well as audio format for ease of communication.**

Keywords*—Sign language recognition, CNN, DenseNet-169, Deep Learning

I. Introduction

Sign Language Recognition (SLR) is an area of interest, as it aims to bridge the communication gap between deaf individuals and hearing individuals. SLR has long been a focus of active research, and recent years have seen considerable breakthroughs in this field as a result of developments in computer vision and machine learning. [1]. American sign language is the most widely used sign language. Deaf and dumb (D&M) individuals can only communicate through sign languages due to a single communication-related disability that prevents them from utilizing spoken languages. Exchanging thoughts and messages through a number of channels, such as speaking, signaling, conduct, and visuals, is the act of communication. D&M individuals utilize a range of hand signals to interact with others. Visual comprehension is used to interpret gesture-based nonverbal communication. The nonverbal method of communication utilized by the dumb and the deaf is sign language [2-3].

With advancements in technology, now it is possible to make communication feasible for all. Recent innovations in machine learning have completely revolutionized the SLR. Supervised learning algorithms such as random forests, decision trees, and support vector machines assisted in development of SLR [4-5]. The process of sign language detection using machine learning involves data acquisition, feature extraction, and classification. In the first stage, sign language data is collected in the form of images or videos. In the second stage, relevant features, such as hand gestures, facial expressions, and body movements, are extracted from the data. In the third stage, machine learning algorithms are used to classify sign language gestures based on the extracted features [6-8].

II. Literature Review

Recent literature utilized CNN and Deep Learning (DL) algorithms for sign recognition systems. Authors in [9] proposed a SLR system based on an advanced LSTM network for indian sign language. They achieved an accuracy of 89.5% for single word recognition, however the system showed poor performance in case of sentence recognition. Rastagoo et al [10] proposed a cascaded model consisting of LSTM, CNN and single short detector (SSD) for hand detection and sign recognition. The proposed model is evaluated on 10000 video poses, its performance is remarkable.

The authors of [11] presented a system for Argentinian sign language (LSA) hand gesture identification. The first of the paper's two major achievements is the creation of a database of handshapes for the Latin American Sign Language (LSA). Second, using a supervised form of self-organizing maps, ProbSom is a technique for assessing photos, extracting descriptors, and then categorizing handshapes. This method is compared to other cutting-edge techniques, such as Support Vector Machines (SVM), Random Forests, and Neural Networks. The ProbSom-based neural classifier achieved an accuracy rate of more than 90% when using the suggested descriptor. The suggested system in [12] has four basic parts: data collection, preprocessing, feature extraction, and classification. Prior to Eigenvector based Feature Extraction and an Eigen value weighted Euclidean distance based Classification Technique, Skin Filtering and histogram matching are performed. Taking into account 24 different alphabets, a 96% recognition rate was found in this study.

ISBN: 979-8-3503-3086-1/23/$31.00©2023 IEEE

Deshpande et al. [13] identify the American Sign Language alphabets that are being signed. The camera captures the hand signing frames, which are subsequently processed by a classifier and a filter to identify the different hand movements being done. The suggested strategy is a first step towards creating a translator for sign language to aid in communication. An HCI system that permits communication with D&M people without the necessity for sign language proficiency is the final result.

Hidden Markov Models (HMM) are employed by [14] to categorize the gestures. The dynamic elements of the gestures are included in this model. The aim is to recognise the two different sorts of gestures—deictic and symbolic. Naive Bayes Classifier is used in [15] as a rapid and effective method for identifying static hand motions. The movements with a static background are taken from each frame of the video, and classified gestures using the K nearest neighbor approach with distance weighting.

There is an urgent need to make communication accessible for all. SLR will play a major role in this. Following are the major contributions:

- Proposed a DenseNet based CNN model for SLR
- Analyzed the system performance using various performance parameters like precision, recall, f1 score, and accuracy.

III. Model Architecture

The model architecture can be broadly categorized as the phases of data gathering, preparation, prediction, and deployment. In this section the deep learning model and system flowchart is discussed.

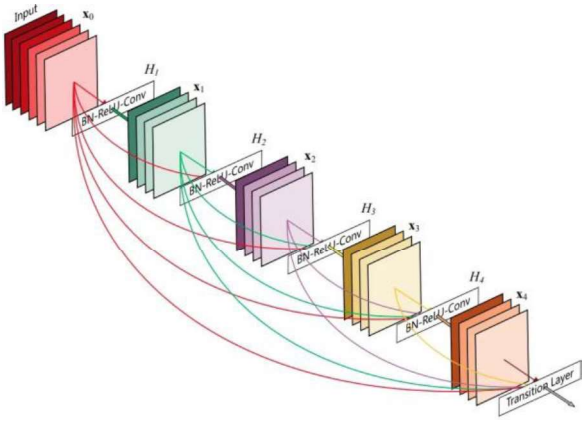

Fig. 1 DenseNet-169 Architecture [16]

A. DenseNet169

CNNs have a variety of technological benefits over fully connected neural networks or other architectures that make them suitable for image processing. From various models included in Keras Applications, such as VGG, ResNet, etc, DenseNet-169 is chosen because it has a low parameter density in comparison to other models and because its architectural design successfully solves the vanish gradient problem. DenseNet (Dense Convolutional Network), provides various advantages over other CNN models such as lower vanishing-gradient, improved feature propagation, feature reuse is encouraged, and the number of parameters is greatly decreased. In Fig. 1 the DesNet-169 architecture is shown. Contrary to other models, such as ResNets, DenseNets concatenates rather than adds the layer's output and input feature maps. Each of the DenseBlocks that make up a DenseNet has a different number of filters, but they all have the same feature map dimensions. These intermediate layers, which are referred to as Transition Layers, use layers of batch normalisation, 1x1 convolution, and 2x2 pooling to handle the downsampling [17-18].

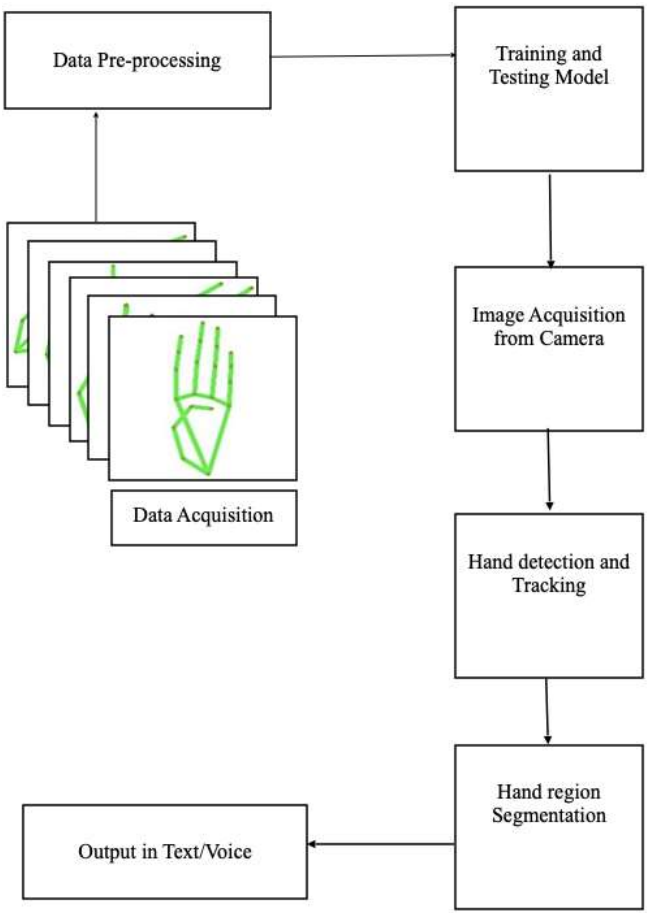

Fig. 2 System Flowchart [15]

B. Working Flowchart

Fig. 2 illustrates the working flowchart of this analysis. The first step is data acquisition. It is performed using a webcam, the user can directly point their hands towards the camera, and it will automatically detect the hand movement. Further, in order to obtain clean data various data pre-processing techniques are employed. Followed bygenerating training and testing datasets using the input data. Later the training and testing datasets will be used to calculate various system parameters. Moreover, using the abovementioned datasets the model performs hand tracking and finally recognises the region of interests using region segmentation. Fin ally after performing all the operations, the model gives the output. Here the output is obtained in text as well as audio, to make communication accessible for all.

IV. Methodology

The suggested approach uses CNN to recognise sign language by videotaping various hand gestures, converting them into frames, and then identifying them. After the hand pixel segmentation, the image is obtained and sent for comparison with the trained model.

A. Data Acquisition & Preprocessing

The computer webcam serves as the input device for vision-based approaches that analyze data from hands and/or fingers. Since only a camera is required to establish a seamless human-computer link, the cost of the Vision Based techniques is reduced. Dealing with the enormous variability in the appearance of the human hand brought on by numerous hand movements, the potential for different skin tones, as well as the various viewpoints, scales, and camera shutter speeds used to capture the scene, is the main challenge in vision-based hand detection. Further, the image quality parameters like contrast, brightness, etc are adjusted for better analysis and prediction. Finally, for incorporating the transfer learning approach, the data labeling is performed.

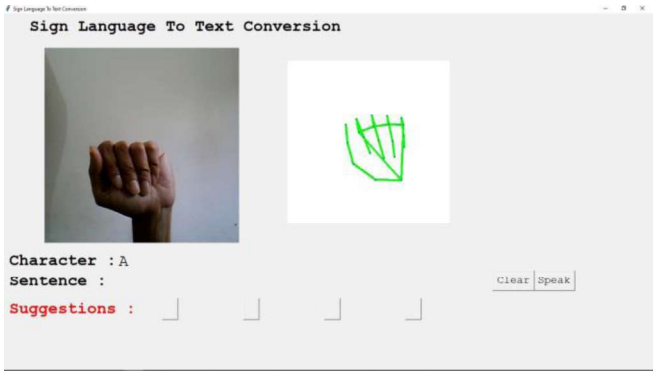

Fig. 3 (a) shows English alphabet with hand gesture

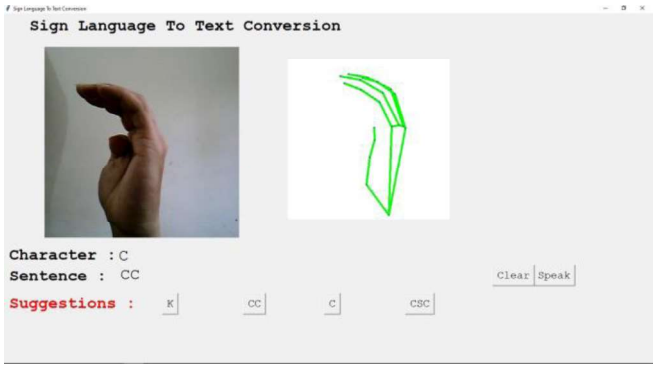

Fig. 3 (b) shows English alphabet with hand gesture

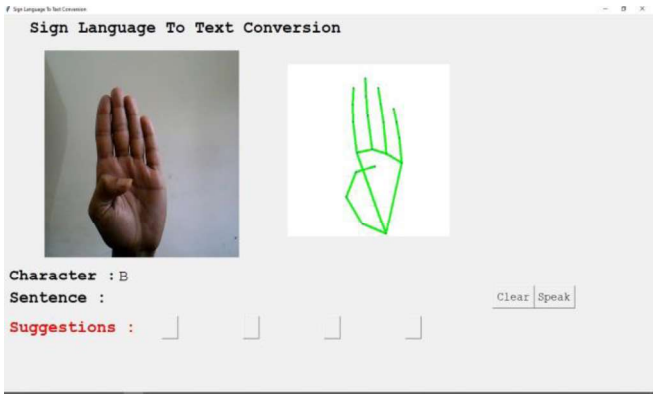

Fig. 3 (a) shows English alphabet with hand gesture

B. Training and Testing Datasets

Using the inputs given by the users, training and testing datasets are created. Training dataset is important for training the DenseNet-169 model on various types of hand gestures and make it capable of recognising the region of interest. Moreover, the testing dataset is important to analyze the system's performance in terms of accuracy, precision, recall and other parameters.

C. Hand Detection & Pattern Recognition

Using the media pipe library, an image processing programme, firstly a hand from a webcam image is selected in order to do hand recognition. Further, the region of interest (RoI), blur is located using a gaussian filter, andthen convert the image to grayscale after removing the hand from the image. The threshold and adaptive threshold techniques are then used to turn the gray image into a binaryimage.

D. Output

After obtaining the RoI, the model predicts the input and gives corresponding output on the computer screen. The output data can be accessed in text as well as audio. The reason for adding the audio output is to increase the ease of communication and make it accessible for all be it blind, deaf or dumb.

V. Result & Discussion

In this section, the effectiveness of the DenseNet169 models is assessed in context with a number of additional factors. The dataset obtained from the webcam is used for this study.

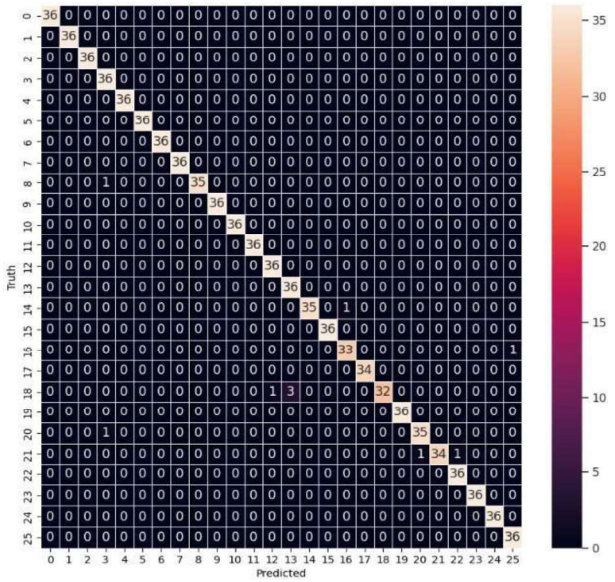

Fig. 4 Confusion Matrix

Fig. 3 (a), (b), (c) shows recognition of different English alphabets respectively using the proposed DenseNet-169 model. In Fig. 4, the confusion matrix for the DenseNet-169 model is displayed. As a result, we were able to calculate performance parameters including precision, accuracy, and f1. Fig. 4 shows the performance metrics for the DenseNet-169 model. The DenseNet-169's accuracy rate was 99%.

Class	Precision	Recall	f1-Score	Support
0	1.00	1.00	1.00	36
1	1.00	1.00	1.00	36
2	1.00	1.00	1.00	36
3	1.00	0.95	0.97	38
4	1.00	1.00	1.00	36
5	1.00	1.00	1.00	36
6	1.00	1.00	1.00	36
7	1.00	1.00	1.00	36
8	0.97	1.00	0.99	35
9	1.00	1.00	1.00	36
10	1.00	1.00	1.00	36
11	1.00	1.00	1.00	36
12	1.00	0.97	0.99	37
13	1.00	0.92	0.96	39
14	0.97	1.00	0.99	35
15	1.00	1.00	1.00	36
16	0.97	0.97	0.97	34
17	1.00	1.00	1.00	34
18	0.89	1.00	0.94	32
19	1.00	1.00	1.00	36
20	0.97	0.97	0.97	36
21	0.94	1.00	0.97	34
22	1.00	0.97	0.99	37
23	1.00	1.00	1. 00	36
24	1.00	1.00	1.00	36
25	1.00	0.97	0.99	37
Accuracy			0.99	932
Macro Avg.	0.99	0.99	0.99	932
Weighted Avg.	0.99	0.99	0.99	932

Table 1 Performance Parameters

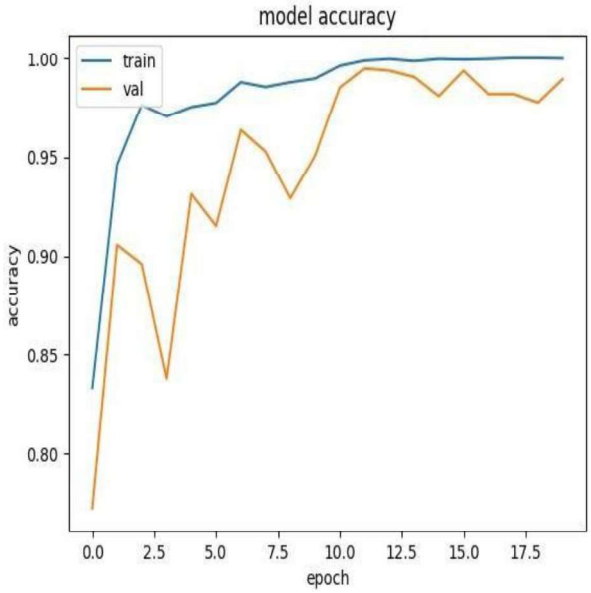

Fig. 5 Accuracy Vs Epochs

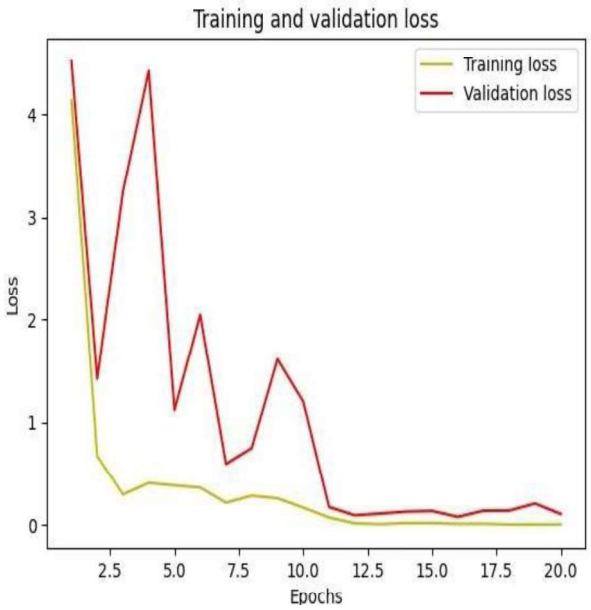

Fig. 6 Loss Vs Epochs plot

Table 1 presents the parameter results obtained using the DenseNet-169 model. It is observed that DenseNet-169 showed excellent results with an accuracy of 99%. Further, Fig. 5 and Fig. 6 displays the accuracy and loss over epoch graphs respectively for the DenseNet-169 model. The graph is presented for both the training and validation data. Up until the first epoch, the accuracy of training data and validation data were equivalent, but after that point, training data accuracy increased more quickly than validation data accuracy.

Conclusion

The SLR method presented in this research is based on the DenseNet-169 CNN. The datasets are obtained from webcams and put through several image pre-processing techniques in order to create a balanced dataset. Using this dataset a training and testing dataset is created for training and evaluating purposes. Additionally, a variety of performance metrics, including accuracy, f1 score, precision, and recall, are used to assess the performance of the proposed model. The proposed DenseNet-169 model demonstrated 99% accuracy. Moreover, the output is obtained in text as well as audio format, so that it can be used by all including blind, deaf, and dumb. To create a workable and affordable communication system for everyone, the proposed model can be combined with contemporary electronic devices. In addition to that, it can be used in public locations like hospitals and schools for making communication accessible for all.

References

[1] Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. Expert Systems with Applications, 164, 113794.

[2] Vaidhya, G. K., & Preetha, C. A. S. (2022). A Comprehensive Study on Sign Language Recognition for Deaf and Dumb people. Journal ofTrends in Computer Science and Smart Technology, 4(3), 163-174.

[3] Upendran, S., & Thamizharasi, A. (2014, July). American Sign Language interpreter system for deaf and dumb individuals. In 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 1477-1481). IEEE.

[4] Ibrahim, N. B., Zayed, H. H., & Selim, M. M. (2020). Advances, challenges and opportunities in continuous sign language recognition.J. Eng. Appl. Sci, 15(5), 1205-1227.

[5] Rastgoo, R., Kiani, K., Escalera, S., & Sabokrou, M. (2021). Sign language production: A review. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3451-3461).

[6] Kakde, M. U., Nakrani, M. G., & Rawate, A. M. (2016). A review paper on sign language recognition system for deaf and dumb people using image processing. International Journal of Engineering Research & Technology (IJERT), 5(03).

[7] Wadhawan, A., & Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. Archives of Computational Methods in Engineering, 28, 785-813.

[8] Nimisha, K. P., & Jacob, A. (2020, July). A brief review of the recent trends in sign language recognition. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 186-190). IEEE.

[9] Mittal, A., Kumar, P., Roy, P. P., Balasubramanian, R., & Chaudhuri, B. B. (2019). A modified LSTM model for continuous sign language recognition using leap motion. IEEE Sensors Journal, 19(16), 7056-7063.

[10] Rastgoo, R., Kiani, K., & Escalera, S. (2020). Video-based isolated hand sign language recognition using a deep cascaded model. Multimedia Tools and Applications, 79, 22965-22987.

[11] Ronchetti, F., Quiroga, F., Estrebou, C. A., & Lanzarini, L. C. (2016). Handshape recognition for argentinian sign language using probsom. Journal of Computer Science & Technology, 16.

[12] Singha, J., & Das, K. (2015). Automatic Indian Sign Language recognition for continuous video sequence. ADBU Journal of Engineering Technology, 2(1).

[13] Deshpande, A., Shriwas, A., Deshmukh, V., & Kale, S. (2023, January). Sign Language Recognition System using CNN. In 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) (pp. 906-911). IEEE.

[14] Fatmi, R., Rashad, S., Integlia, R., & Hutchison, G. (2017). American Sign Language Recognition using Hidden Markov Models and Wearable Motion Sensors. Trans. Mach. Learn. Data Min., 10(2), 41-55.

[15] Amrutha, K., & Prabu, P. (2021, February). ML based sign language recognition system. In 2021 International Conference on Innovative Trends in Information Technology (ICITIIT) (pp. 1-6). IEEE.

[16] Huang, G. (2017). Dense connected convolutional neural networks. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

[17] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv preprint arXiv: 151203385.

[18] Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Fractalnet: Ultra-deep neural networks without residuals. arXiv preprint arXiv:1605.07648.