Hand Gesture Recognition Using Machine Learning with Convolutional Neural Network (CNN)

Shubham A. Sangale[1], Narayan B. Kirtane[2], Avinash A. Bhatane[3], Madhuri S. Jagtap[4], R. M. Samant[5]
Students, Information Technology, NBN Sinhgad School of Engineering, Ambegaon BK., Pune[1,2,3,4]
Guide, Information Technology, NBN Sinhgad School of Engineering, Ambegaon BK., Pune[5]

Abstract: *Many deaf and hard-of-hearing people rely heavily on sign languages as a means of communication. Sign languages are the native languages of the Deaf community and provide full access to communication. Despite the fact that it is an effective mode of communication, communicating with speech impaired people remains a barrier for those who do not understand sign language. The purpose of this paper is to create a Web application that will convert hand gesture to English in the form of text. hence facilitating sign language communication. The web application uses the computer's webcam to collect visual data, which is then pre-processed using a combinational method before being recognized via template matching. This project tried different machine learning algorithms including SVM, RNN and CNN (Convolutional Neural Network). With the accuracy of 91% CNN algorithm has proven to be the most accurate classification tool.*

General Terms: Hearing-Impaired People, Computer Vision, Hand Gesture Recognition.

Keywords: Sign Language, ASL, Hearing Disability, Convolutional Neural Network (CNN), Computer Vision, Machine Learning, Gesture Recognition, Sign Language Recognition, etc.

I. INTRODUCTION

All humans engage in activities such as reading books, singing, and playing on a daily basis. Now this project focused on singing, let's talk about deaf or dumb people, those who can't speak or hear. This group of people is known as the disabled community. As a result, we discovered a significant communication gap between the normal and disabled communities.

There is a boy who lives in residency. The boy was unable to express himself due to his disability (deaf& dumb). As a result, communicating with him can be challenging at times. His family was also attempting to comprehend him, but there was something missing, and he was not able to express himself properly. The only reason is that he is deaf and dumb boy. After reaching the boy and his family it is observed that there are some daily routine work and actions to communicate with his family. According to that daily routine work and actions, 40 different words were collected and used this project.

Language is unquestionably important to human interaction and has existed since the dawn of civilization. It is a medium through which individuals communicate in order to express themselves and comprehend real-world concepts. It is so deeply ingrained in daily lives that people frequently take it for granted and overlook its significance. Unfortunately, in today's fast-paced society, those with hearing impairments are frequently neglected and excluded. They must strive to communicate themselves to others who are different from them, to bring up their thoughts, to speak their opinions, and to express themselves. Although sign language is a means of communication for deaf individuals, it has no significance when communicated to non-sign language users. As a result, the communication gap has widened. Motive of proposing a sign language recognition technology to prevent this from happening. It will be a fantastic tool for persons with hearing impairments to convey their thoughts, as well as a great way for non-sign language users to grasp what the latter is saying.

DOI: 10.48175/IJARSCT-5275

Many countries have their own set of sign motions and interpretations. For instance, an alphabet in Korean sign language will not mean the same thing as in Indian sign language. While this highlights diversity, it also pinpoints the complexity of sign languages. Deep learning must be well versed with the gestures so that it can get a decent accuracy. In a proposed system, made-up hand gestures are used as dataset. The image is then transformed into words. Hand movement is observed as part of the detecting process. The method returns a text output, which helps to reduce the communication gap between deaf-mute and people.

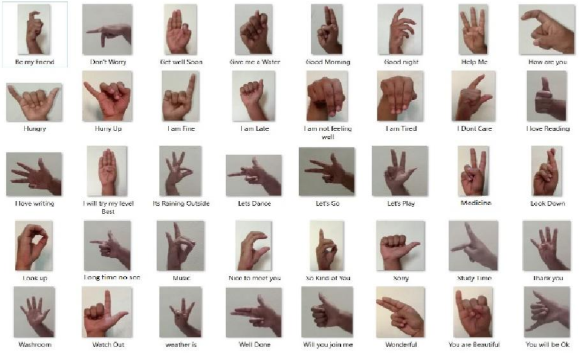

Figure 1: Different Hand Gestures and Their Meanings

This are the daily routine words of deaf and dumb people. This hand gestures used to denote the words to communicate with disable community.

II. LITERATURE REVIEW

Literature review of this proposed system reveals that many attempts have been made to solve sign identification in videos and photos using various methodologies and algorithms. Project work was extended by using made-up sign language. In this project, every thought of deaf or dumb person has different hand gestures.

Siming He et al. [1] proposed a system using a 40-word dataset and 10,000 sign language graphics. Faster R-CNN with an incorporated RPN module is utilized to locate the hand regions in the video frame. In terms of accuracy, it enhances performance.

When compared to single stage target detection algorithms like YOLO, detection and template classification can be done at a faster rate. When compared to Fast-RCNN, the detection accuracy of Faster R-CNN improves from 89.0 percent to 91.7 percent in the paper. For the language image sequences, a 3D CNN is employed for feature extraction, and a sign-language recognition framework comprising of long- and short-time memory (LSTM) coding and decoding networks is created. The paper combines the hand locating network, 3D CNN feature extraction network, and LSTM encoding and decoding to develop an extraction technique for RGB sign language picture or video recognition in practical scenarios. In the common vocabulary dataset, this paper received a 99 percent recognition rate.

Let's look at the work of Rekha, J et al. [2], who used the YCbCr skin model to detect and fragment the skin region of hand gestures. The picture characteristics are retrieved and categorized by Multi class SVM, DTW, and non-linear KNN using Principal Curvature based Region Detector. For training, a dataset of 23 Indian Sign Language static alphabet signs was employed, and 25 videos were used for testing.

M. Geetha and U. C. Manjusha et al. [3] used 50 examples of each letter and digit in a vision-based recognition of Indian Sign Language characters and numerals using B-Spine approximations in their paper. The sign gesture's region of interest is analyzed and the boundary is removed. The acquired boundary is then converted to a B-spline curve utilizing the Maximum Curvature Points (MCPs) as Control points. A number of smoothing processes are applied to the B-spline curve in order to extract features. The photos are classified using a support vector machine, which has a 90.00 percent accuracy.

For Bengali hand gesture recognition, Muhammad Aminur Rahaman et al. [4] employs a unique approach. To detect the hand in each frame, the system employs cascaded classifiers. It records hand gestures using the HIS color model's Hue and Saturation values. The K-Nearest Neighbours Classifier is then used to classify the photos.

P. Gajalakshmi et al. [5] used Support Vector Machine and Error Correcting Output Codes for American Sign Language recognition.

In 1996, Triesch and Malsburg created an ASL hand gesture recognition system with the goal of performing accurate gesture recognition even in images with complex backgrounds. They have eliminated handwriting in this system.

III.PROPOSED METHODOLOGY

Convolutional Neural Network (CNN)

CNN is a Deep Learning approach that is based on the Multi- Layer Perceptron and is used to handle two-dimensional data. CNN is divided into two steps during the process: Feature Learning and Classification. The act of transforming information from an image into numbers that show an image is known as feature learning. This portion has two layers: a convolutional layer and a max pooling layer (optional), as well as an activation function to alter values to a specific range. Numbers in the feature map will be translated into vector shapes, and the data will be categorised by type. In this test, the variety of architectures is used; the difference between them is the number of layers and their location.

Convolution Layer:

Convolution is a data processing technique that extracts several properties from a single set of data. In the first stage, it extracts low-level features like edges and corners. Then, at a higher level, upper-level layers extract functionality. For the 3D convolution process in CNNs. The input is N x N x D in size, and it's convolved with H kernels that are each k x k x D in size. One output feature is generated when one input is convolutional with one kernel, and H features are produced when H kernels are convolutional individually. Starting at the top-left corner of the input, each kernel is shifted from left to right. If a kernel enters the top-right corner, it is relocated down one element before being moved from left to right one element at a time. The technique is repeated until the kernel reaches the screen's bottom-right corner.

Convolution is a mathematical technique that accepts two inputs, such as an image matrix and a filter or kernel. The image matrix is a digital representation of picture pixels, and the filter is another matrix used to process it. It can process any aspect of the image because the kernel is significantly smaller than the image. Apply a filter to this layer's image matrix. To execute convolution, this filter matrix is combined with the image matrix shown in Figure. Depending on the features to be eliminated, any number of convolution layers can be added. The number of filters, the structure of each filter, the input shape, and the image form and resolution are the four arguments for the convolution function. The fourth input specifies the triggering function to be utilised. Which neuron can fire next is determined by the activation mechanism?

Pooling Layer:

CNN has another construction block in the form of a pooling layer. The primary goal of pooling is to reduce the size of the image matrix. Its purpose is to make computing the big picture matrix resulting from the convolution operation easier. Each feature map is treated separately by the pooling layer. To reduce the complexity of processing these matrices, pooling must be performed on the output matrix after the convolution procedure. The fundamental goal of a pooling process is to down sample the matrix in order to reduce its size.

Flattening Layer:

Flattening is a technique for converting a pooled function map into a single column that can be transferred to a fully linked layer. The process of flattening data into a one-dimensional sequence so that it can be passed on to the next layer is known as flattening. Flatten the contributions of the convolutional layers to create a single long function vector. It's attached to the final categorization model, which is a fully linked sheet.

Fully Connected Layer:

The entry to the fully connected layer is the final pooling or convolutional layer's contribution, which is flattened before being fed into the completely connected layer. Each neuron in this layer corresponds to a weight, which is chosen at random. This calculation is thus performed at each layer as g(wx+b). Here x represents the input vector, w represents the weight, b represents the bias, and g represents the activation function. This process is repeated for each layer. The CNN network's completely linked component uses backpropagation to calculate the most reliable weights. Weights are assigned to each neuron to prioritise the most appropriate mark. Finally, the neurons "vote" for one of the marks, and the assignment decision is based on the outcome of that vote. After going through the Entirely Connected Layer connected layers, the final layer is used to calculate the probability of the input being in a specific class.

3.1 Architecture

The image in Fig. 2 is taken in three dimensions by CNN (width, height, depth). As a result, neurons are grouped in this way. Each layer of CNN accepts 3D input and converts it to 3D neuron activity output. If it's an RGB image, the depth will be 3, and the image's height and width will be the dimensions. CNN does a good job using images. Convolution, pooling layers, and fully connected networks make up CNN. The next chapter goes over each tier of CNN.

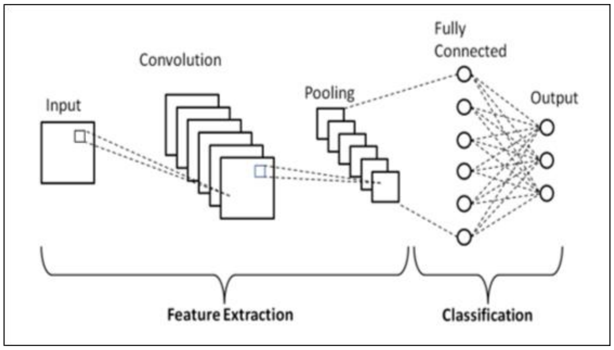

Figure 2: CNN Architecture

3.2 System Architecture

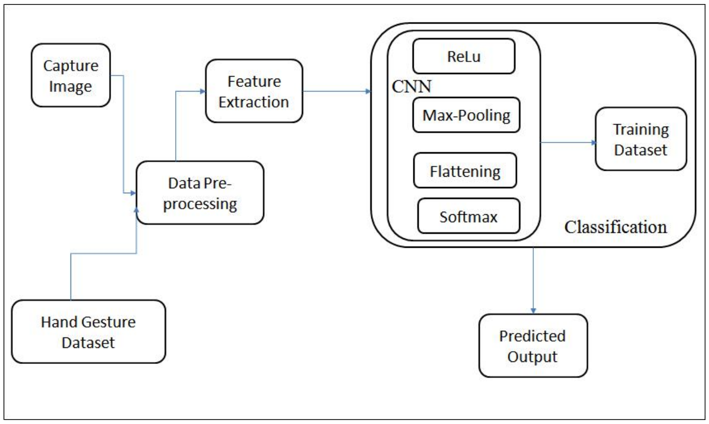

Figure 3: System Architecture

DOI: 10.48175/IJARSCT-5275

3.3 Training the Convolutional Neural Network (CNN)

Adjusting the weights of individual neurons to extract the relevant information from images is one of the most difficult aspects of constructing CNNs. "Training" the neural network refers to the act of altering these weights.

The CNN starts with random weights at the beginning. During training, the developers submit a big collection of photos labelled with their matching classifications to the neural network (cat, dog, horse, etc.). Each image is processed with random values, and the outcome is compared to the image's correct label.

If the network's output does not match the label which is likely at the start of the training process it adjusts the weights of its neurons to get its output closer to the correct response the next time it views the same image. Back propagation is a technique used to make the corrections (or backprop). Back propagation, in essence, enhances the tuning process by allowing the network to choose which units to alter rather than making random modifications.

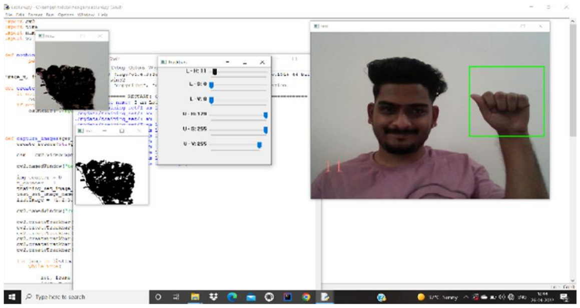

Figure 4: Training the CNN Algorithm

IV. RESULTS & DISCUSSION

The task of executing real-time object detection with quick inference while retaining a base level of accuracy is known as real-time object detection. This real time gesture detection is used to detect the hand gesture for the made-up sign language. In this project some specific signs are used that denote some specific emergency and daily routine needs.

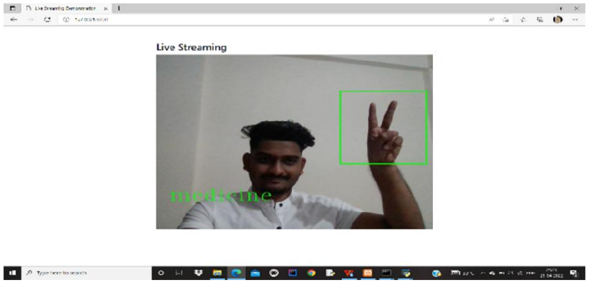

Figure 5: Live View of System

Above image shows the live view of system. At the time of change in hand gesture system automatically tries and detects meaning of that hand gesture.

Support Vector Machines

On all of the feature vectors, multiclass SVMs were used. For all feature vectors, the results of linear kernel and fourfold Cross Validated accuracies are presented. The confusion matrices in the following sections correspond to the various strategies that were explored using linear kernel Multi Class SVMs. This method was found to have the second highest accuracy of 72 %.

Recurrent Neural Network

RNN (Recurrent Neural Network) is also used because it is also a type of deep learning algorithm which can be used for gesture detections. This method was found to have the third highest accuracy of 60 %.

Convolutional Neural Network

The training results of the Convolutional Neural Network show an accuracy of up to 100%, indicating superior performance in all factors. Validation results, on the other hand, have an average accuracy of around 91 %. For hand gesture recognition, the CNN algorithm has proven to be the most accurate classification tool.

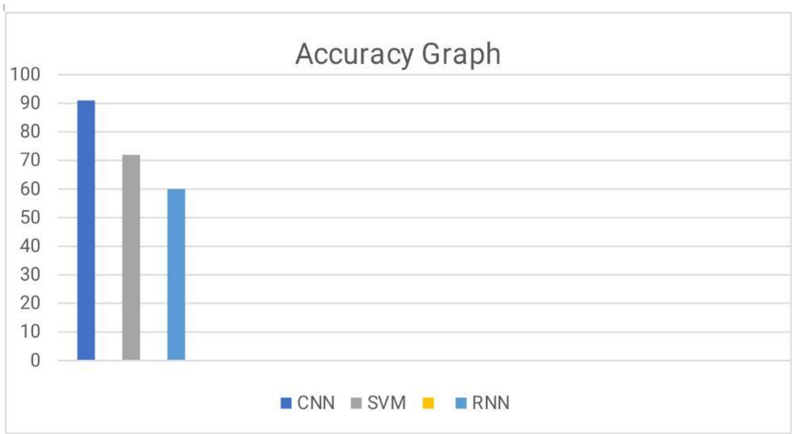

Figure 6: Accuracy Comparison of CNN, SVM, RNN

V. CONCLUSION

Under various lighting conditions and speeds, the suggested system successfully predicts the signs of sign and certain frequent words. The photographs are accurately masking by providing a range of values that can recognize a human hand dynamically. For picture training and classification, the proposed method employs CNN. More informative elements from the photos are finely extracted and used for classification and training. For each sign, a total of 4000 static photos are utilized in the training process to ensure accuracy. The system can recognize a total of 40 words, including alphabets. Thus, this is a user-friendly system that can be easily accessed by all the deaf and dumb people.

REFERENCES

[1] K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4896-4899, doi: 10.1109/BigData.2018.8622141.

[2] Jaoa Carriera, A. Z. (2018). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (pp. 4724-4733). IEEE. Honolulu.

[3] Jing-Hao Sun, Ting-Ting Ji, Shu-Bin Zhang, Jia-Kui Yang, Guang-Rong Ji "Research on the Hand Gesture Recognition Based on Deep Learning",07 February 2019.

[4] Chaudhuri, Arindam and Mandaviya, Krupa and Badelia, Pratixa and Ghosh, Soumya K and others. (2017) “Optical Character Recognition System. In Optical Character Recognition Systems for Different Languages with Soft ComputingSpringer: 941.

[5] Li, Haixiang and Yang, Ran and Chen, Xiaohui. (2017) “License plate detection using convolutional neural network. 3rd IEEE International Conference on Computer and Communications (ICCC), IEEE:17361740.

[6] Geethu G Nath and Arun C S, "Real Time Sign Language Interpreter," 2017 International Conference on Electrical, Instrumentation, and Communication Engineering (ICEICE2017).

[7] Sanmukh Kaur, Anuranjana (2018), “Electronic Device Control Using Hand Gesture Recognition System for Differently Abled”, 8th International Conference on Cloud Computing, Data Science & Engineering.

[8] Saransh Sharma, Samyak Jain, Khushboo (2019), “A Static Hand Gesture and Face Recognition System for Blind People”, 6th International Conference on Signal Processing and Integrated Networks.

[9] Anush Ananthakumar (2018), “Efficient Face and Gesture Recognition for Time Sensitive Application”, IEEE Southwest Symposium on Image Analysis and Interpretation.

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