Evaluation of Different Techniques for Sign Language Recognition

Bhuvana Raisinghani[1], Manasee Parulekar[2], Shubhranshu Arya[3], Varun Bhat[4], Prof Manisha Chattopadhyay[5]

[1-4]Student, Dept. of Electronics and Telecommunication Engineering, Vivekanand Education Society's Institute of Technology, Maharashtra, India
[5]Assistant Professor, Dept. of Electronics and Telecommunication Engineering, Vivekanand Education Society's Institute of Technology, Maharashtra, India

---

Abstract - *Sign language is a way of communication using hand gestures. It is a communication tool used by people who are unable to hear or speak. While being unique and independent of spoken languages, sign language is an important tool for seamless communication between people. Sign language is geography specific and hence there are multiple conventions for it. In this paper we consider ASL or American Sign Language as our basis and compare various models for sign language translation using Machine Learning algorithms. We review papers based on sign language translation and formulate our findings in the form of advantages and disadvantages. We then select a framework which best suits all the criteria for a good sign language translation system and explain its benefits over other machine learning algorithms.*

Key Words: Communication tool, Seamless communication, American Sign Language, Machine Learning, Comparison

1. INTRODUCTION

Privileged are those who have the means to communicate efficiently without any natural hurdles. Unfortunately, this is not the case for a lot of people in the world. According to a study by World Federation of the Deaf (WFD), around 72 million people in the world use sign languages for communication. Sign languages play a massive role in day-to-day conversations and functioning of the deaf and mute community. Sign languages have minor changes in different geographic locations and hence we cannot term one sign language of a region as universal. Machine learning is a type of Artificial Intelligence that allows systems to self-educate based on the content provided to it as training data. The system, after educating itself through various algorithms, is proficient to predict necessary targets for evaluation of various trends. This is a growing technology with growing prospects of usage in almost every sector of technology and finance.

1.1 Application of Machine Learning in sign language translation:

Machine Learning can be successfully utilised in sign language translation. Machine learning has various algorithms that can be used to predict a sign and translate it into English or any other language as per training. This is the need of the hour as through social media and video conferencing platforms we are constantly trying to bring the world closer one step at a time. In such an environment, non-inclusion of sign language creates a barrier between the deaf and mute and the normally abled people. Integration of Machine Learning to predict the signs of the deaf and mute requires a user to have a functional web-camera. This is because the person using sign language will have to perform the signs with the reference frame of the web-camera so that through the real-time system, the Machine Learning algorithms can detect the signs through the camera and predict the words on the screen. This leads to a seamless communication between people who know sign language conventions and people who don't.

2. LITERATURE REVIEW

The following paper [1] explains how hand gestures and facial expressions are recognized using the skin segmentation feature of OpenCV. The raw videos taken in a dynamic background is given as an input to the real-time sign language recognition system wherein high-intensity noises from the video frames are eliminated. The blurred image is then obtained by performing a convolution operation with a low-pass box filter and the images are converted from BGR color spaces to HSV color spaces. Binary images are obtained at the end of the preprocessing.

Morphological transformations are operated on these binary images based on the shape of the image. Contours work very well on binary images hence contours are used for detecting the object. To group similar data items, FCM assigns membership to each data point based on its proximity to the cluster centers. Several factors are taken into account and It is well known that FCM yields the best results, but it requires more computation time than the others. Also, traditional algorithms have difficulty in handling high-dimensional datasets. Therefore, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) merging will allow the system to capture both temporal and spatial features.

In [2] Vishwas S, et al. have proposed that recognitions of various hand gestures and representation as corresponding words can be achieved with the Decision Tree classifier algorithm. The real-time data in the form of video, acquired from the webcam, is processed by the pose estimation library from tensor-flow. It is administered to map the skeleton of

the person enacting the hand gesture. The 17 points are captured across the body and are converted into 3D coordinates per point. All these coordinates are stored in a CSV format. Each gesture is thereby distinguished by a set of 51 overall values. The labeling is done in the CSV file accordingly. These markings are then sent to the tree classifier to make predictions. An increased number of frames will ameliorate the efficiency of the model. The approach is taking into consideration a wide range of factors including shoulders and facial coordinates for better results. The ineluctable shortcoming associated with the model is that it requires training data from a huge number of people since the distance between two points in the skeleton varies largely. This would result in inaccuracy. Also, the intricate hand gestures won't be captured and displayed properly since points on fingers, thumb, and palm are not taken into consideration.

In [3] J. L Raheja, et al. implemented a hand gesture recognition system for Indian sign language where the fixed number of frames are obtained from the real-time video input. These frames are later passed through the skin filters to uproot Hue moments, features of the palm and fingertips, and the tracked trajectory, this is done by converting frames into HSV color space. The noise from the images is also attenuated for smoother images. These features are added to the empty vector created before. If the feature vector length matches that of the number of frames, these gestures are then compared to the database that has been defined before. If the frames and features match then the sign is recognized.

In this paper [4], the authors used a real-time model for ISL gesture recognition, based on the incoming image data from the Microsoft Kinect RGB-D camera. Further adding computer vision techniques like 3D construction and affine transformation, pixels of the depth and the RGB camera were mapped together as there was no one to one mapping. RGB-D pairs were trained simultaneously using double-channel CNN based architecture. The images are then re-sized into a size of 120 into 120 keeping the aspect ratio constant. They are passed through a series of convolutional layers, ReLu layers and max-pooling layers where they are reduced to 90 feature maps of size 7x7. The images are then passed onto two fully-connected layers with 1024 and 144 neurons respectively. Finally, SoftMax activation classifies the gestures in the last layer. The combined RGB + Depth based CNN model attained a training accuracy of 98.81% and the testing set accuracy reached 99.6%. Even though the adopted model provides the best result, there was a good deal of computational heaviness observed in 3D CNNs. For the dynamic dataset, artificial data synthesis was not adopted due to limitations in the computational resources.

The following paper [5] explains how Sign Language is recognized using LibSVM, a library from Support Vector Machines. The raw videos taken in a dynamic background is given as an input to the real-time sign language recognition system wherein noises from the video frames are eliminated and the video is the processed for further detailing. The Skin Color Sampling helps in obtaining tighter range for the hue and saturation of human in the footage and helps in noise reduction. Then, the face is detected and removed using the Voila and Jones Algorithm. Then the hand is isolated using the HSV Thresholding using the range sampled earlier. The largest contour is then detected and filled to obtain a clear image of the gesture. This image is then post processed, as finger tips are determined using a range of set angles between the fingers and palm. To identify the category of the gesture, we compute the distance moved by the Center of Gravity (CoG) of the hand in subsequent frames, let it be 'G' as represented in the flow diagram. This distance G is compared with at-most two thresholds, T' and T'', at two separate stages of the flow. For static gestures, we use Zernike Moments to identify the orientation of the hand, and for dynamic we involve tracking of CoG or a particular fingertip's pathing. Once the required features have been extracted, the SVM is used to classify the gesture to its corresponding alphabet. The speech recognition was implemented in Android using the speech engine Sphinx 4. The flow of the Android Application comprises of a decision node which decides whether the speech input corresponds to a letter involving a static or a dynamic gesture for which an image is displayed or a video is played respectively.

In this paper [6] convolutional neural networks have been implemented to overcome the drawbacks of a normal neural network. A normal neural network is not capable of processing multiple images with varied information that have to be retrieved in a short span of time, as each neuron is individually connected to each layer of a neural network. The Convolutional neural network processes multiple images with varying values, most of the existing pixels of these images carry unrequired information so it reduces the dimension of the images repeatedly as it is processed in layers through filters, this process is called convolving. Images of gestures in American sign language have been used for the dataset. It contains 24 static gestures from the letter A to Y, excluding dynamic gestures of the letter J. Over 100 images have been used and 20 from these have been used specifically for testing. Like the CNN model the inception model too stacks its inputs but unlike CNN model, here the layers are stacked parallel to each other. Then the outputs are linked and fed as inputs to the next block. Inception blocks are repeated loops of the same operation; these architectures enable faster processing of large collections of data using minimum computing power. Hence cutting off the requirement of large processing and computing power usually needed by such processes making it easier for devices with lesser or limited computing power

3. MACHINE LEARNING ALGORITHMS

Accuracy - The accuracy of a machine learning model is based on the input, or training, data and determines which model is the best in identifying relationships and patterns between variables. A good accuracy score signifies that the

probability of unerring outputs of the total outputs is satisfactory.

For training, we can take into consideration four machine learning models to compare which algorithm gives the best result. The four machine learning models for an efficient system are as follows:

• Logistic Regression – This is an algorithm which is used for the classification problems. It can be sub-classed into a predictive analysis algorithm which is based on the concept of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function.

Training Model	Accuracy Score
Logistic Regression	0.922

• Ridge Classifier - This is a model tuning method that is utilized to work with any data that contains substantial amount of multicollinearity. This method performs L2 regularization. The least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values for multicollinearity.

Training Model	Accuracy Score
Ridge Classifier	0.932

• Random Forest Classifier - This is a classifier algorithm that contains a number of decision trees on several subsets of the dataset in question and takes the average of all the decision trees to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Training Model	Accuracy Score
Random Forest Classifier	0.94

• Gradient Boosting Classifier - Gradient boosting classifiers are a group of machine learning algorithms that assimilate many poor learning models together to create one strong predictive model. Decision trees are usually used when doing gradient boosting.

Training Model	Accuracy Score
Gradient Boosting Classifier	0.88

4. CONCLUSIONS

Through implementation of various Machine Learning algorithms, we have found out that the Ridge Classifier and the Random Forest Classifier are providing better accuracy and precision. These tuning methods use regularization and feature selections to analyze data with greater efficiency. Machine Learning is a felicitous approach for bridging the gap in communications and acts as a modern technique to predictions and evaluation of trends across a wide range of sign languages. This novel proposition can be utilized to help the whole of a deaf and mute community.

5. RESULTS:

The following are the results of the predicted hand sign of our system:

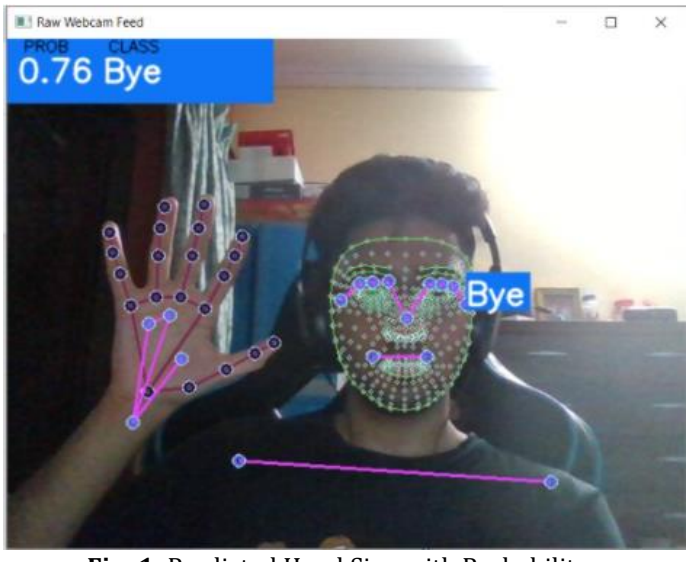

Fig -1: Predicted Hand Sign with Probability

REFERENCES

[1] Muthu Mariappan H, Dr Gomathi V "Real – Time Recognition of Indian Sign Language" Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019)

[2] Vishwas S, Hemanth Gowda M, Vivek Chandra H N, Tannvi "Sign Language Translator using Machine Learning" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 4 (2018)

[3] J.L. Raheja, A. Mishra, A. Choudhary, "Indian Sign Language Recognition using SVM" Pattern Recognition and Image Analysis volume 26, pages434–441 (2016)

[4] Neel Kamal Bhagat, Vishnusai Y, Rathna G N "Indian Sign Language Gesture Recognition using Image Processing and Deep Learning" 2019 Digital Image Computing: Techniques and Applications (DICTA)

[5] Anup Kumar, Karun Thankachan and Mevin M. Dominic "Sign Language Recognition" 3rd InCI Conf. on Recent Advances in Information Technology I RAIT-20161

[6] Aditya Das, Shantanu Gawde, Khyati Suratwala, Dr Dhananjay Kalbande "Sign Language Recognition using Deep Learning on Custom Processed Static Gesture Images" 2018 International Conference on Smart City and Emerging Technology (ICSCET)