JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# SIGN LANGUAGE CONVERSION TO TEXT AND SPEECH

**Medhini Prabhakar [1] Prasad Hundekar[1] Sai Deepthi B P[1] Shivam Tiwari[1]**
Vinutha M S[2]
[1]Department of Computer Science and Engineering, Dr. Ambedkar Institute of Technology, Bangalore, 560056
[2]Assistant Professor, Department of Computer Science and Engineering, Dr. AIT

**Abstract— This system presents a novel approach for translation of sign action analysis, recognition and generating a text description in English language and then conversion of the generated text to speech. In training set there were 26 Indian Sign Language Alphabet image samples used whereas testing captures hand gestures from the live feed and predicts the class label based on several trained models like CNN (Convolutional Neural Networks), FRCNN(Faster-Convolutional Neural Networks), YOLO(You Only Look Once) and Media Pipe. Finally, the text description will be generated in English language and converted to speech. The average computation timeis bit more than expected due unavailability of high GPU hardware but has acceptable recognition rate in case of FRCNN model. When it comes to CNN model, the recognition of hand gestures is fast enough for real world applications but there is a compromise in accuracy of identification. YOLO model recognizes the sign language with a good accuracy but is not satisfactory in case of speed as it is taking more time when live feed of hand gestures is captured and converted. Though YOLO model works inefficiently for real time conversion, it performs greatly when already captured hand gestures are fed as input to the model. Media Pipe model of sign language conversion checks all the requirements of our project which include great accuracy as well as real time conversion of hand gestures to text to speech in real time without any delay.**

*Keywords— Convolutional Neural Networks(CNNs), FRCNN(Faster-CNN), YOLO(You Only Look Once) , Media Pipe*

## I. INTRODUCTION

The only form of communication for deaf and mute people—mostly illiterates—is sign language. However, it is still difficult to interact with regular people without the aid of a human interpreter because most members of the general public are not eager to learn this sign language. The deaf and hard of hearing become isolated as a result. Nevertheless, the development of technology makes it possible to overcome the obstacle and close the communication distance.

Various sign languages are used around the globe. There are around 300 different sign languages in use around the globe. This is so because individuals from various ethnic groups naturally created sign languages. Maybe there isn't a common sign language in India. Different regions of India have their own dialects and lexical differences in Indian Sign Language. However, new initiatives to standardize Indian Sign Language have been made (ISL).It is possible to train the machine to recognize gestures and translate them into text and voice. To facilitate communication between deaf-mute and vocal persons, the algorithm effectively and accurately categorizes hand gestures. Additionally, the identified sign's gesture name is spoken and displayed. system helps the blind to navigate independently using real time object detection and identification.

This is a software-based project which uses Media Pipe framework to detect hand gestures. There are a couple of other algorithms like FRCNN, CNN and YOLO

which were used to achieve the objective of the project, but each model had its own cons. Finally, Media Pipe model satisfies all requirements of the project and provides output as expected.

The project first captures hand gestures through a web camera and processes the captured frame by undergoing preprocessing and segmentation and hence identifies hand gesture shown by the signer in real time. The recognized sign is also converted to a voice form which also increases the use cases in which this project can be used.

This is a software-based project which uses MediaPipe framework to detect hand gestures. There are a couple of other algorithms like FRCNN, CNN and YOLO which were used to achieve the objective of the project but each model had its own cons. Finally, MediaPipe model satisfies all requirements of the project and provides output as expected.

## II.RELATED WORKS

PAPER 1:

An Efficient Approach for Interpretation of Indian Sign Language using MachineLearning.

*Abstract*:

This paper focus on the most accurate translation of spoken English words into Indian Sign Language gestures as well as standard Indian Sign Language gestures into English. For this,various neural network classifiers are created, and their effectiveness in recognising gestures is assessed. The suggested ISL interpretation system accomplishes two key tasks: I Converting gestures from text to gestures from speech. On the pre-processed photos, feature extraction is performed. This entails turning the raw data (pictures) into numerical characteristics so that the classification algorithm can process the information. The information contained in the original data is kept even though the image is translated to numerical form. Convolutional neural networks and other machine learning algorithms are fed the retrieved image features as input. Support Vector Machine (SVM) and Recurrent Neural Network (RNN).Google Speech Recognition API and PyAudio are used to convert speech to text. The Keras framework was used to model and create a convolutional neural network. Python library The classifier model was trained using about 30,240 photos, which represents 60% of the dataset's total image count. The classifier was trained using various epoch counts.A test accuracy of 88.89% was found to be the average.Thirty-two,240 photos were utilised to train the model of a classifier. The classifier was trained using various epoch counts. Around 82.3 percent total testing accuracy at its highest level was attained.

PAPER 2:

Hand Gesture Recognition For Automated Speech Generation

*Abstract*:

The field of gesture recognition has developed recently, and new tools, gadgets, models, and algorithms have come into existence. Despite numerous advancements in the aforementioned technology, humans still feel comfortable making motions with their hands alone. This research focuses on a system that runs on a mobile computing device and gives us the technology for automated translation of the Indian Sign Language system into Speech in the English Language, enabling bidirectional communication between people with vocal impairments and the general public. Since it operates on the Gesture Recognition concept, this system can be utilised in the near future for communication between individuals who do not comprehend sign language. The system uses an internal mobile camera to recognise and capture gestures. The captured gestures are then analysed using algorithms such the HSV model-(Skin Color Detection), LargeBlob Detection, Flood Fill, and Contour Extraction. The standard alphabets (A-Z) and numeric values can be represented using one-handed signs that the system can identify (0-9). The results of this system's gesture processing and voice analysis are very accurate, reliable, and efficient.

PAPER 3:

Indian Sign Language Recognition –A Survey.

*Abstract*:

This paper focuses on various methods for understanding Indian Sign Language. With various tools and algorithms applied on the Indian sign language recognition system, a review of hand gesture recognition methods for sign language recognition is reviewed, along with difficulties and potential directions for future research. The acquisition of the signer's image, or the individual communicating through sign language, can to be captured with a camera. The acquisition process can be started manually. To record the signer's characteristics and gestures, a camera sensor is required. Scaling an image is utilised to cut down on the computational work required for processing images before skin detection. This procedure produces a binary image in which the hand is defined by pixels that are coloured white and all other pixels are black. Each pixel of the image is classified throughout this processing as either being a part of human skin or not. Extracting features lessens the precision without sacrificing computing time. There are many features that can be measured, including hand shape, hand orientation, textures, contour, motion, distance, centre of gravity, etc.used to identify sign language. Principal Component Analysis (PCA) benefits from the decreased dimensionality. Though PCA is highly susceptible to the scaling and

rotation and translation of the image, so before using PCA, the image needs to be normalised. The vision-based technique is user-friendly because a signer does not need to wear the bulky gloves. The way a gesture appears when it is being identified relies on a number of factors, including the camera's position and the signer's proximity to it. These techniques have been utilised to keep the real-time performance's accuracy and computing complexity in check.

## III.METHODOLOGY

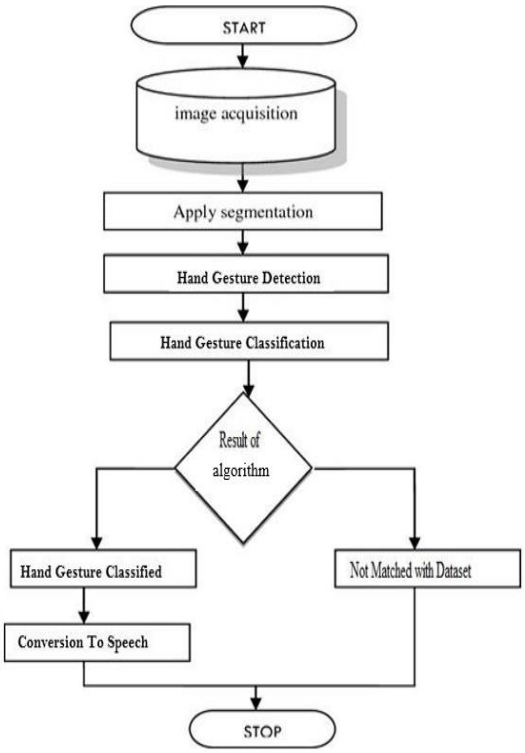


Fig: Methodology of Sign Language Conversion To Text And Speech Model

The model architecture is presented here. This includes the way image is acquired, segmented, the flow of hand gesture classification. This will show how image is captured and converted to speech as the output. The interaction of image data with the model is shown in the architecture diagram above.

It uses the user's video and translates the video into the frames (Multiple images). Each frame will then go through preprocessing, which will be discussed in the following section, prior to extracting features. The web camera's video is still being continually recorded and divided into a number of frames. A frame is a term used to describe every single image. Video framing is the technique of deleting specific frames from a given video by making use of video characteristics like frame rate. Cropping, filtering, brightness and contrast adjustments, among other things, are all included in image preprocessing. Image enhancing, image cropping, and image segmentation techniques are employed in this procedure. The format of captured images is RGB. Therefore, the first step is to convert RGB photos to binary images, and then crop the image to get rid of any unnecessary parts. Additionally, improvements can now be made in a specific, chosen region. Edge detection techniques are used in image segmentation to locate the border of cropped images, which are then employed in feature extraction techniques. Classification data is used to assign corresponding level with respect to groups with homogeneous characteristics, with the aim of discriminating multiple objects from each other within the image.

Classification involved in two Steps: -

Training step: In this step, with use of the training samples matrix, the method calculates the parameters of a probability distribution, considering features are conditionally independent given the class.

Testing step: For any untested sample, the method finds the posterior probability of that sample belonging to each class. The method then classifies the test sample according to the largest posterior probability.

At last, when the classification process is completed the equivalent grammatical text description will be generated with the help of the class labels which are assigned during the training phase. Finally, using Google API Conversion of Text to Speech will be done.

## IV.IMPLEMENTATION

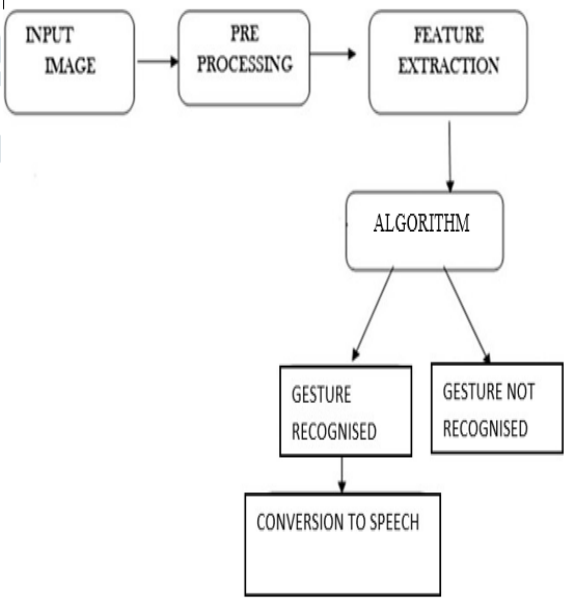


Fig: Flowchart

Step 1: The first stage is to segment the skin part from the image, as the remaining part can be regarded anoise w.r.t the character classification problem

Step 2: The second stage is to extract relevant features from the skin segmented images which can provesignificant for the next stage i.e., learning and classification.

Step 3: The third stage as mentioned above is to use the extracted features as input into the algorithm for training and then finally use the trained models for classification.

Models used for implementation:

1. FRCNN for Indian Sign Language

We used a data set of double-handed gesture photos in Indian Sign Language that comprised both alphabet and numbers. Data was given to the FRCNN model. The Fast R-CNN detector employs an algorithm similar to Edge Boxes to produce region proposals, just like the R-CNN detector does. The Fast R-CNN detector processes the complete image as opposed to the R-CNN detector, which shrinks and resizes region proposals. Fast R-CNN pools CNN features corresponding to each area proposal, whereas an R-CNN detector must categorise each region. Because computations for overlapping regions are shared in the Fast R-CNN detector, it is more effective than R-CNN.

2. CNN for American Sign Language

The English alphabet and numbers in the American Sign Language data set that we downloaded from Kaggle were each represented by a single hand. A Convolutional Neural Network is a Deep Learning method that can take in an input image, give various elements and objects in the image significance and be able to distinguish between them. Comparatively speaking, a ConvNet requires substantially less pre-processing than other classification techniques. ConvNets have the capacity to learn these filters and properties, whereas in primitive techniques filters are hand-engineered. The Convolution Operation's goal is to take the input image's high-level characteristics, such edges, and extract them.

3. YOLO for American Sign Language

To meet the project objective and purpose and get better results, we implemented the YOLO model for sign language conversion. You Only Look Once is known by the acronym YOLO. This algorithm identifies and finds different things in a picture. The class probabilities of the discovered photos are provided by the object identification process in YOLO, which is carried out as a regression problem. CNN is used by the YOLO method to recognise items instantly. The approach just needs one forward propagation through a neural network to detect objects, as the name would imply. This indicates that a single algorithm run is used to perform prediction throughout the full image. Multiple class probabilities and bounding boxes are simultaneously predicted using the CNN. This project makes use of the YOLOv3 version.

4. Media Pipe for Customized Sign Language

In this model, the classes are defined in a custom manner for each hand gesture. Several phrases which people use for daily conversations are included as a part of training data set which is fed to the Media pipe framework. For these jobs, MediaPipe, an open-source framework created especially for complicated perception pipelines utilising accelerated inference (e.g., GPU or CPU), currently provides quick and precise, yet distinct, solutions. It is a particularly challenging task that necessitates simultaneous inference of numerous, dependent neural networks to combine them all in real-time into a semantically consistent end-to-end solution. Today, MediaPipe Holistic offers a fresh, cutting-edge human pose topology that opens up new use cases as a solution to this problem.

## V. OBSERVATIONS

1. ISL model using FRCNN shows good accuracy but does not achieve the mark in terms of speed.
2. ASL model using CNN happens to be acceptable for real time applications in terms of speed but is not satisfactory in terms of accuracy.
3. YOLO model of ASL ticks the mark for both accuracy and speedy performance. Nonetheless, the model shows good efficiency only when the input is fed to the model manually but lays back when it comes to taking input in the form of live feed.
4. Media Pipe model achieves all expectations and objectives of our project which is sign language recognition and conversion to text and speech from in real time.

## VI. RESULTS

Real-time image capture from the camera is given to the algorithm model for processing. The Media Pipe concept is utilised by the python code to identify and categorise the objects. It will outline the detected area with border boxes and display the object's category index. A text file will be used to keep the category index of the objects that were detected. The class name and class id of the discovered object make up the category index. Following hand gesture identification, the text representation of the hand gesture is transformed to speech. The user may easily transport this system because it is portable.

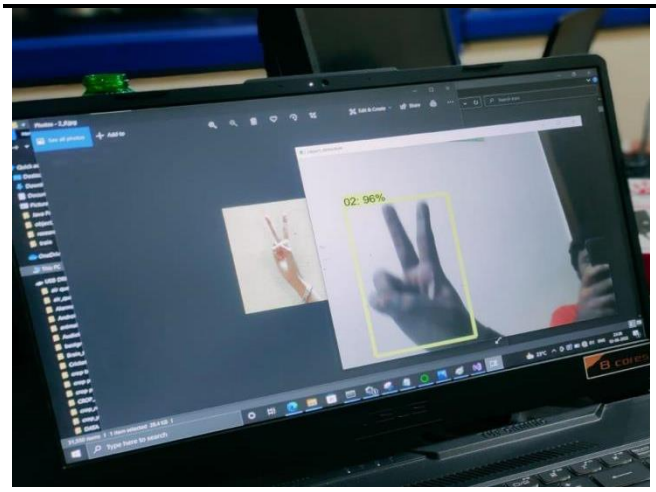


Fig: Detecting the number 2

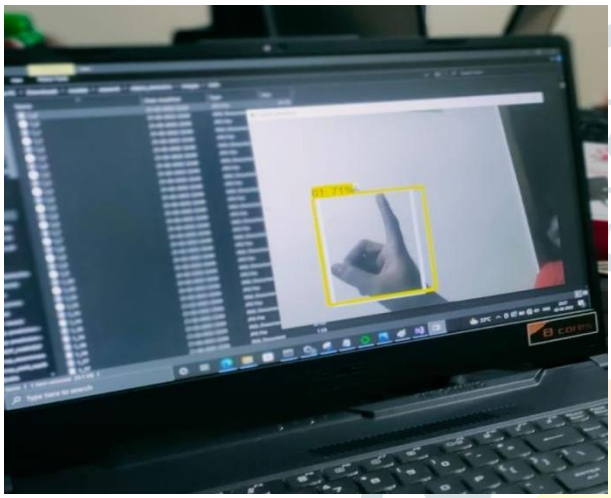

Fig: Detecting the number 1

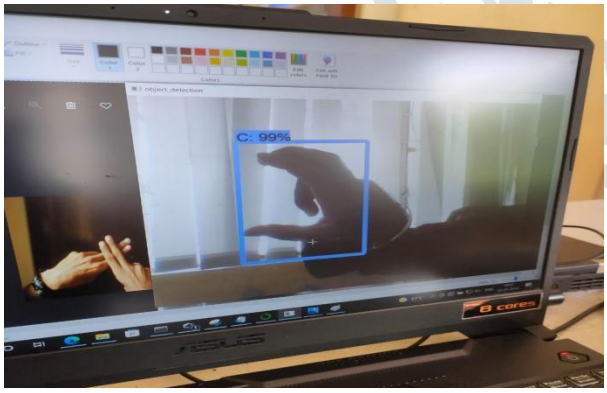


Fig: Detecting Sign C

## VIII. CONCLUSION AND FUTURE WORKS

This project uses four different algorithms to recognize sign language. The FRCNN algorithm crosses the mark of good accuracy but fails to achieve performance in terms of speed required for usage in real world. The second algorithm which was used to implement this project was CNN for ASL. This model was performing well in terms of speed but remained behind in terms of accuracy which was again not hitting the brief of our project objective. YOLO algorithm was used in this project with an aim of achieving good accuracy as well as speed in the same instance which was lacking in the first two models. Though the conjunction of good accuracy as well as speed was achieved, the model was not collecting input in real time. Keeping these hinderances into consideration, the project was finally implemented using Media Pipe algorithm. The algorithm classifies alphabet in sign language efficiently with a good number of accuracy and also the identified hand gesture is converted to speech for a better result so that it be used for communication not only among deaf - mute and vocals, but also can be applicable for visually impaired people. Regarding this problem, proposed system is developed to solve communication problems for vocally disabled people and there by encouraging all people to make better interactions and hence not make them feel isolated.

To improve accuracy and applicability of this Project work can be further extended on Server Based Systems which will be implemented to improve the coverage. High Range Cameras can be used to getthe better sign detection. By interconnecting such systems with a central computer will help in accumulating the data and hence the performance of the whole network is benefited with this exchangeof data since it will help to train the algorithm better.

## REFERENCES

[1] https://gilberttanner.com/blog/tensorflow-object-detection-with-tensorflow-2-creating-a-custom-model/

[2] https://learnopencv.com/introduction-to-mediapipe/

[3]https://towardsdatascience.com/yolo-object-detection-with-opencv-and-python-21e50ac599e9

[4] Indian Sign Language Recognition System for Deaf People " Int. J.Adv. - A. Thorat, V. Satpute, A. Nehe, T.Atre Y.Ngargoje

[5] Machine Learning Techniques for Indian Sign Language Recognition, International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC-2017) - Kusumika Krori Dutta, Sunny Arokia Swamy Bellary

[6] A. Sharmila Konwar, B. Sagarika Borah, C. Dr.T.Tuithung, “An American Sign Language Detection System using HSV Color Model and Edge Detection,” International Conference on Communication and Signal Processing, pp. 743-746, Melmaruvathur, India, April 3-5, 2014

[7] Anuja v. Nair, Bindu V, "A Review on Indian Sign Language Recognition," International Journal of computer Applications, pp.33-38, July 2013.

[8] A Novel Feature Extraction for American Sign Language Recognition Using Webcam - Ariya Thongtawee, Onamon Pinsanoh, Yuttana Kitjaidure

[9] Dhivyasri S, Krishnaa Hari K B, Akash M, Sona M, Divyapriya S, Dr. Krishnaveni V An Efficient Approach for Interpretation of Indian Sign Language using Machine Learning , 2021 3rd International Conference on Signal Processing and Communication | 13 – 14 May 2021 | Coimbatore

[10] J. R. Pansare, S. H. Gawande and M. Ingle, "Real-Time Static Hand Gesture Recognition for American Sign Language in Complex Background," Journal of Signal and Information Processing, No. 3. pp. 364-367