iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: I Month of publication: January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48695

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Real Time Indian Sign Language Recognition using Deep LSTM Networks

Jainam Rambhia[1], Manan Doshi[2], Rashi Lodha[3], Stevina Correia[4]
[1, 2, 3, 4]Department of Information Technology, Dwarkadas J Sanghvi College of Engineering

Abstract:** On our planet, people with speech and hearing disabilities are part of society. Communication becomes difficult when it is necessary to interact with survivors and the general public. In some races, people with disabilities practice different sign languages for communication. For people with speech and hearing disabilities, sign language is a basic means of communication in everyday life. However, a large portion of our community is unaware of the sign languages they practice, so bringing them into the mainstream is an incredible challenge. Today, computer vision-based solutions are still well received for helping the general public understand sign language. Many analysts are trying out one of his computer vision-based sign language recognition solutions, Recognition of Hand Gesture. Lately it's been a popular area of research in various vernacular languages being used across the world. Through this research work, we propose a solution to this problem using Keypoint identification and Neural Network architecture for real time sign language recognition. This architecture uses a Long Short Term Memory architecture (LSTM) giving the best accuracy of 91.5% in prediction of words as well as phrases of Indian Sign Language.

***Keywords:** Computer Vision, Sign Language, Deep Learning, Pose Estimation, Artificial Intelligence, Mediapipe*

I. INTRODUCTION

According to the WHO survey on deaf and hearing loss, approximately 1.5 billion people globally live with hearing loss, which is about 19 percent of the world's population. Also, by 2030 it states that around 2.5 billion people will be in this curve of hearing impairment. Most of these people are from the younger generation, with an age group of 12 to 35 years old. Furthermore, WHO also states that, with this impairment, the overall annual cost of unaddressed hearing loss globally is about 980 billion dollars, which consists of costs of educational support, healthcare costs, productivity loss, and costs in society. While it is also clear that the deaf and the deafened encounter many greater difficulties than the general public because it may affect one year or sometimes both the years in certain cases. Communication is crucial in our daily lives because it allows us to pass information from one person to another. Deaf and dumb persons, on the other hand, find it extremely difficult to converse with people with no impairments. One of the most efficient ways in order to bridge the gap amongst us is through sign language. Normal individuals that are people with no impairs, on the other hand, are not aware of sign language. As a result, it leaves us with one and the only option: translate sign language to text and voice. In India, there are different regional languages spoken across the entire country, depending on the tribes, states, regions and areas. As a result, there exists no mainstream Indian Sign Language. The fact that Indian sign language may communicate with both single and double hands furthermore complicates the challenge of detecting it. Sign language comprises of three main factors: body language, hand gestures, and facial expressions each of it having it's own importance. However, out of all of them hand gestures play a very crucial role when it comes to communicating with a set of actions. Thus, the majority of research has been focused on decoding and identifying those hand actions. This can be done in two ways:

1) Sensor-based detection
2) Vision-based / Image-based detection

Sensor based methods mainly suggests using physical devices such as hand gloves and sensors. These are used to extract sign language with the motion of the hand. The intricate movements of the hands and fingers are captured very accurately using the sensor-glove combination. Some of it also takes into consideration the area of the body where the action is being performed (Gestural Space) which is then classified using computer algorithms to determine the sign language. However, the sensor-based method is complex and costly, and for a large-scale implementation, it may not prove to be beneficial. On the other hand, vision-based identification is the more preferred way to solve this problem. Considering the vision- based method using a web-camera as a study, it can be very easily implemented on any platform, such as mobile phones for daily use with no extra costs involved for the user.

The ease of use also increases with this method of implementation. For example, with no requirement for gloves or sensors, and by only capturing gestures, identification of Sign Language and transmission of the gestured communication using text and sound becomes way more easier.

Detection of sign language is one of the topics that has been discovered for more than 3 decades now, and with the evolving technologies, it seems that we are not really far from real-time sign language detection. Real-time sign language detection do poses different challenges for example taking into consideration the computation power for large-scale implementation, getting into the details of time frames , etc. The deep-learning algorithms devised for the detection are computationally heavy processes. Thus this involves taking the problem to the next level by taking the help of IT infrastructure so as to solve the problem. Considering only vision-based systems for our research, it can again be further subdivided into two sections:

a) Static vision-based detection

b) Dynamic vision-based detection (Real-Time)

In static vision based detection, input images are fed to the system and the output derived gives us the detection of alphabets and numbers in terms of text and sound. This was one of the first systems to investigate sign-language detection, at the same time the processing power required in this method is really less for computations. However, when we look into dynamic vision-based detection systems, continuous video input is provided to the system, which in turn does the task of dividing the input into frames. After getting these image frames they are then further mapped, processed, and classified using machine learning algorithms to detect sign language. Finally the output is then given in terms of text and speech. Our proposed system contains pose estimation techniques to detect and track hand movements in each frame which is then further used to distinguish signs. This method achieves state of the art performance on Indian Sign Language dataset. An With the current advancement in Deep Learning techniques a lot of research is being done in order to bridge the gap between people using sign language and the ones not using it. However none of the models till now have been trained that well for it to be used commercially. Sign language can be classified in both manual and non manual body gestures and physical devices can also be used in order to detect the motion like using a Microsoft Kinect and 5DT Sensor Gloves (6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info exclusion (DSAI 2015) Virtual Sign – A Real Time Bidirectional Translator of Portuguese Sign Language) A comprehensive survey has revealed the problems that arise due to finger spelling signs within video streams causing decreased precision in the model [2]. The advancement in Indian Sign Language has not been in parallel to the American Sign Language. Hossein et al. [3] used CNN architecture to identify native Bengali Sign Language's alphabets and gives an accuracy of 99.86%. However no work has been performed on phrases or words in particular. In [5] the authors have used Sign Language Graph Convolution Network (SL-GCN) and SSTCN to get a detailed analysis of the skeleton features then using CNN's for RGB and RGBD modalities (SAM-SLR approach) giving accuracy of 98% in each of them. Speaking with those who have hearing loss is really difficult. Since deaf and mute people utilize hand gestures to communicate, normal people have difficulty understanding their language from the signals they make. Systems that can identify various indications and provide information to common people are thus necessary. The deaf use sign language as a form of communication. Deaf or vocally impaired people in India utilize Indian Sign Language for communication. With various tools and algorithms implemented on the Indian sign language recognition system, a review of hand gesture recognition methods for sign language recognition is continually updating, along with obstacles and future research direction. The amount of people with this difficulty is increasing exponentially compared to the count of expert Sign Language Tutors. According to World Health Organization, 63 million persons are either completely or partially deaf in India among which 50 million are children. To solve this problem various solutions have come-forth. For the current development stages Indian sign language recognition has been done with the help of recognition of alphabets and letters. Latest researches in 2019, demonstrated the use of Fuzzy Clustering [6] and CNN. The former method first preprocessed images by removing high intensity noises from the image, smoothing and blurring the image by a low-pass box filter operation. For better extraction images were converted from RGB to HSV format. Morphological transformations were performed which comprises of dilation and erosion. Appropriate background noise is removed by median blurring techniques and contours are found that joins all the points. Feature extraction plays a prominent role before the process of applying fuzzy clustering which is used to determine and identify the action with an accuracy of 75%. The latter research uses CNN with color space model and thresholding where primary task is to convert RGB color to YCbCr scheme . After which, Otsu's segmentation method is used before applying DI-CNN (Double Input Convolution Neural Network) over the obtained regions with a maximum accuracy of 99.96%. One of the Deep Learning Method is used to identify static gestures with depth variant using Microsoft Kinect RGB-D camera with static model achieving and accuracy of 98.81% and dynamic model achieving an accuracy of 99.08%. Over these developments, more advanced machine learning algorithms were used to identify phrases and words in Indian Sign Language Recognition in Real-Time.

One of the faster and more advanced Computer Vision Algorithm YOLO was used for real-time Indian Sign Language Detection. YOLOV3 is an advanced and faster version of the CNN family of algorithms. Initially, video data is fed to frame extraction which is further annotated that is creating target regions or bounding boxes around desired objects in the image. Furthermore, YOLOV3 also known as Darknet-53 is used to train and identify gestures like "Change", "Correct", "Good", "Sorry" etc. Samples are used to train and then tested on real-time with precision ranging from 92.3% to 93.6% with video and images inputs respectively.

II. WORKING METHODOLOGY

A. Dataset Acquisition

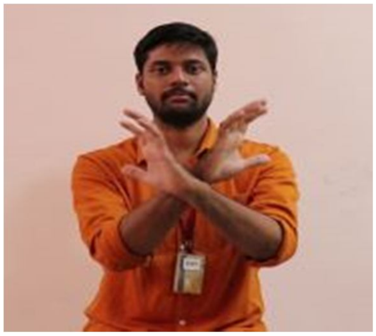

Fig. 1 Sample frame from video of a gesture

The camera's input method is the dataset acquisition phase. Later on, the camera's collected frames will be analysed and used as input by the neural network. Microsoft Kinect cameras have been used by numerous researchers. It is capable of simultaneously providing depth and colour video streams. As stated in [7] background segmentation can be carried out easily using Kinect for sign language recognition. Our research was focussed on real world use case of this technology which will make use of normal cameras with resolution of most cameras being 480p and hence we chose a dataset consisting of a large vocabulary of 700 fully annotated videos, 18663 sentence level frames and 1036 word level images for 100 spoken language sentences performed by 7 different signers [1]. The completely labelled videos assisted us to build an end to end framework for converting spoken language sentences into sign language. We have performed analysis on sentence phrases and singular words dataset for predicting the sign language.

B. Processing Data

The capacity of the neural network to interpret the flow of the sign in each frame was impeded by the fact that several of those images overlapped and had high similarity scores. We calculated the similarity score using the Structural Similarity Index. Since the scores given by the similarity index are between -1 and +1, we kept the threshold of 0.75 for similarity. Any image that crossed this threshold was removed from the training set as it depicted high similarity. Hence we were able to produce a series of images that accurately depicted the movements of a sign from its beginning to its end.

III.PROPOSED APPROACH

A. Hand Tracking

Land-marking as many as 21 key-points on each hands is the key component of the system as these same 42 keypoint locations i.e 21 from each hand is then being fed to the Neural Network for detection of Sign Language. We have used Mediapipe Detection library which is provided by Google for extracting hand landmarks locations. To elaborate this, hand tracking pipeline consists of two models:

1) A palm detector that operates on a full input image and locates palm by creating bounding boxes and
2) A hand landmark model that operates on the cropped hand bounding box provided by the palm detector. The model has three outputs:
3) 21 hand landmarks consisting of x coordinates, y coordinates , and it's relative depth.
4) A hand flag indicating the probability of hand presence in the input image.
5) A binary classification of handedness, e.g. left or right hand.

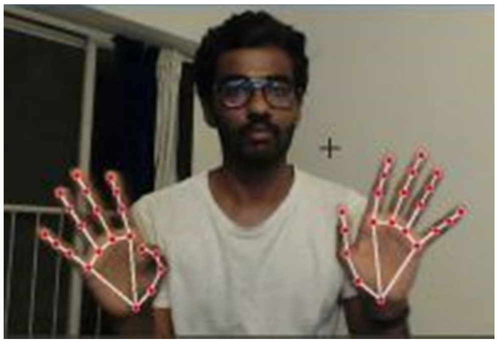

Fig. 2 Hand Keypoints

Each sign language of the phrases takes approximately three seconds to complete. The video is cut into 30 frames and the hand tracking model tracks the movement of hands in all the frames. It forms an array of (30,126) where 30 are the number of frames and 126 has been derived from the keypoints on hands. Each hand has 21 keypoints and each keypoint is located using the x, y and z axis. Hence 21x2x3 = 126.

B. LSTM Architecture

Deep LSTM networks are bulit by stacking multiple LSTM layers. Deep LSTM networks have proved to be more accurate than traditional LSTM models. As observed in [4] with two layers of LSTM RNNs, the performance improves but still it is not very good. The LSTM RNN with five layers approaches the performance of the best model. Training an seven layer LSTM is hard as the model starts converging after a day of training. The input to the Sequential model will be the keypoints of both the hands extracted using the mediapipe library. Indian sign language is more focused on the movement of hands and not dependent on the facial features hence the keypoints of face have not been used as input for the model. The sequential model helps us to store data of each frame in a sequential manner. The parameters of LSTM such as units, activation function, dropout and bias regularizer were tested with different values. The optimum accuracy was achieved on testing when the sequential model created contained three LSTM layers and three Dense layers. The activation function used for all layers except the last Dense layer is Rectified Linear Unit. As the data is categorical the activation function used for the last dense layer is softmax function. For each of the 50 signs on which the model has been trained on, the model gives the output as the likelihood of each phrase. The phrase which has the highest likelihood score is the predicted sign. For real time recognition, we have used computer vision library for capturing real time video. The video is then converted into frames which are then sent for keypoint extraction.

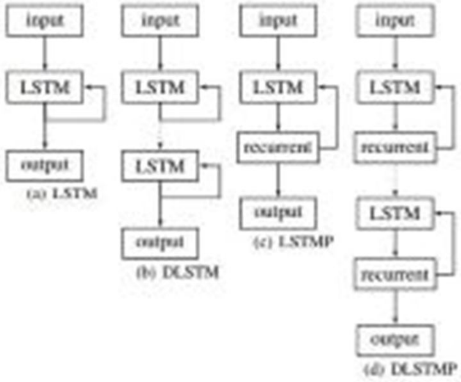

Fig. 3. Deep LSTM Networks

IV. CONCLUSION

We started with sign language image predictions. Following a 95% accuracy with image recognition, we moved on to real-time video identification of Indian sign language. Using hyperparameter optimization we were able to train the dataset on the right number of LSTM layers without overfitting the model and obtained an accuracy of 91.5%. Compared

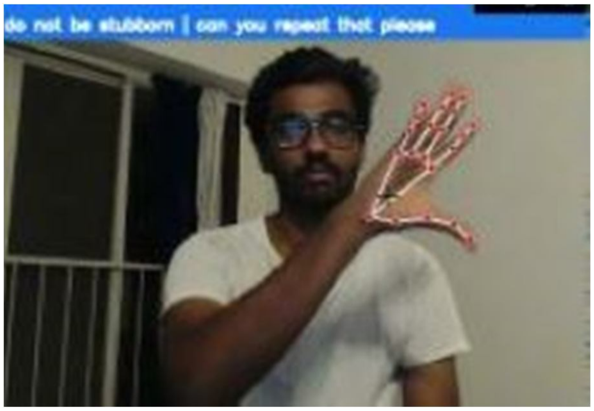

Fig. 4 Real Time Video Prediction

to the previously deployed box detection approaches which gave poor accuracy (80%) in real time video recognition, the adoption of the keypoint detection model produced improved accuracy. We want to use this technology to improve communication for the deaf and mute, and we want to learn more about its practical applications. We believe that our findings will stimulate and assist further study into the identification of sign language.

REFERENCES

[1] R, Elakkiya; B, Natarajan (2021), “ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition”, Mendeley Data, V1, doi: 10.17632/kcmpdxky7p.1

[2] El-Sayed M. El-Alfy, Hamzah Luqman, A comprehensive survey and taxonomy of sign language research, Engineering Applications of Artificial Intelligence, Volume 114, 2022, 105198, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2022.105198.

[3] M. J. Hossein and M. Sabbir Ejaz, ”Recognition of Bengali Sign Language using Novel Deep Convolutional Neural Network,” 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, pp. 1-5, doi: 10.1109/STI50764.2020.9350418

[4] H. Sak, A. W. Senior, F. Beaufays, ‘Long short-term memory recurrent neural network architectures for large scale acoustic modeling’, 2014.

[5] Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y.. (2021). Skeleton Aware Multi-modal Sign Language Recognition.

[6] H. Muthu Mariappan and V. Gomathi, ”Real-Time Recognition of Indian Sign Language,” 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862125

[7] M. S. Subburaj, S. Murugavalli, Survey on sign language recognition in context of vision-based and deep learning, Measurement: Sensors, Volume 23, 2022, 100385, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2022.100385.