International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | June 2022 www.irjet.net p-ISSN: 2395-0072

VIDEO BASED SIGN LANGUAGE RECOGNITION USING CNN-LSTM

Sai Bharath Padigala[1], Gogineni Hrushikesh Madhav[2], Saranu Kishore Kumar[3], Dr. Narayanamoorthy M[4]

[1,2,3]BTech CSE, Vellore Institute of Technology, Vellore.
[4]Associate Professor, Dept of IoT, VIT, Vellore.

Abstract - *Sign Language Recognition, SLR, is an important and promising tool for assisting hearing-impaired persons in communicating. The goal of this work is to develop an efficient solution to the sign language recognition problem. This study uses deep neural networks to create a sign language recognition framework that directly transcribes films of SL signals to English and speech. Previous techniques for SL recognition have often used hidden Markov models which is not very efficient for sequence modeling that involves temporal analysis. Convolutional Neural Networks, CNN, are used for spatial characteristics, and Recurrent Neural Networks, RNN, are used for temporal analysis.*

Key Words: Sign Language(SL), Gesture Recognition, Convolutional Neural Network(CNN), Recurrent Neural Network(RNN), Transfer Learning, Spatial Features, Temporal Features.

1. INTRODUCTION

According to the World Health Organization(WHO), more than five percent of humans suffer from hearing impairment. Sign language(SL) is a language that is used to communicate with the help of hands and expression by making hand movements, and it is often used to communicate with hearing or speech impaired people. sign language is often regarded as one of the best grammatically structured languages which is very easy to learn. Because of this reason, SL recognition is a good study subject for creating algorithms to handle problems like Gesture Recognition, Sequence Modelling, and user interface design, and it has received a lot of interest in multimedia and computer vision. Take input from the user in the form of a video where the user wears green and red coloured gloves on their left and right hand respectively. Use Convolutional Neural Networks (CNN) to get the spatial features of the hand to feed them to the RNN. Use Recurrent Neural Networks (RNN) to analyze the hand movements to recognize the sign in the video given as input. Convert the recognized sign from RNN to text and speech.

2. RELATED WORK

There have been many methods used to tackle Gesture recognition problems. Some used RNN with Long Short Term Memory(LSTM) to train the model, choosing angles between the bones of fingers as training features to train the neural network[4]. To acquire the hand features they used a Leap Motion Controller (LMC) sensor with the support of infrared light, which can track the hand movements with high precision but using the LMC is not a feasible solution. [3] To determine the boundary between each sign, Reinforcement learning has been used to divide the signs in the video. Weakly supervised Learning has been used for training. The first stage's output is utilized as the second stage's input. [2] Used CNN to detect and extract features like the hand, and face details and also both hands have been taken into consideration. used Bi-LSTM for temporal analysis which is not required as it is not very necessary for sign language because future signs will not decide the probability of the present sign. [1] Inspired by the Automatic speech recognition, they used Transition action modeling for segmentation of signs in the video, so they can be classified just like in isolated video format, but transitions can be very indefinite and subtle, so finding transitions can be ambiguous using the proposed approach. So inaccurate segmentation leads to a significant loss.

2.1 Feature Extraction:

Earlier we have seen [17]-[19] using hardware-supported devices to extract the spatial features, but of late many new deep learning approaches has been came to light with high accuracy and efficiency for gesture recognition [11],[20],[21]. It shows that CNN is one of the best approaches to extract the features than using Hardware devices which is costly and not feasible. for Video-based sign language recognition, there are two problems to deal with, one is spatial features and another one is temporal analysis. Many popular architectures of CNN like ResNet, Inception v3, AlexNet, GoogleNet etc., [22]-[25] are very good at extracting features from any given image as they have been trained on images to a greater extent with very large datasets. 2D-CNN and 3D-CNN are becoming popular techniques to use for video classification problems like gesture analysis[27] [28], determining the person [29] and

describing and captioning the video tasks [30]. 3d-CNN has also performed well both for space and time series analysis with high efficiency [20], [31] -[34]. To do the temporal analysis, Recurrent Neural Networks have been widely used. [21] used a heaped version of convolution to distinguish the boundaries between signs of continuous Sign Language. Some used the combination of various deep learning concepts [10], [26] for temporal feature analysis.

2.2Sequential Analysis

For sequential data, Recurrent Neural Network has been using a lot these days, Earlier Hidden Markov Model used to be the one used for this kind of activities. but now RNN has been very impressive with the data that has time series attached to it. two types of RNN, LSTM[37] and GRU[38] are being used in many time series related tasks like speech recognition[35], Forecasting problems, Prediction problems that have sequential figures, neural machine translation[36], sign language translation[40][12] etc., because they help in solving the problems that basic RNN cell does. basic RNN does not have the memory to withhold the memory of previous data from that is from long back. An architecture, that has encryption that will get the subject of the signs and decryption that will give the final result, has also done very good with SL recognition. Many hybrid versions of different architectures of RNN, CNN, 3D-CNN, and CTC have been used to make these SL detection problems and sequential modelling related tasks [6],[10],[16].

2.3 Limitations

Usually, we use image classification for sign language but it won't work for all the signs, so there is no scalability for new signs that take motion to perform. And for video classification, some hardware is used to capture the hand spatial features to find the sign, but it requires extra hardware to use and is also not very accurate as well.

3. PROPOSED METHOD

COLOUR SEGMENTATION

Colour segmentation is used to segment-specific colours from any given frame or image. The required colour RGB values range will be considered and pixels out of that region will be blacked out. Pixels in the range will be extracted. We used colour segmentation to extract fists of the hands. The user wears coloured gloves for both hands with different colours. We extract red and green colours from the frames and merge them into one frame where we can extract only hands from the image or frame. Figure-1 shows the user wearing gloves and performing a sign facing the camera. In Figure-2, we can see that the hands of the user have been extracted using colour segmentation and are represented in grayscale image format.

Fig-1 User Wearing coloured gloves and enacting sign 'APPEAR'

Fig-2 Extracted hands from the above figure

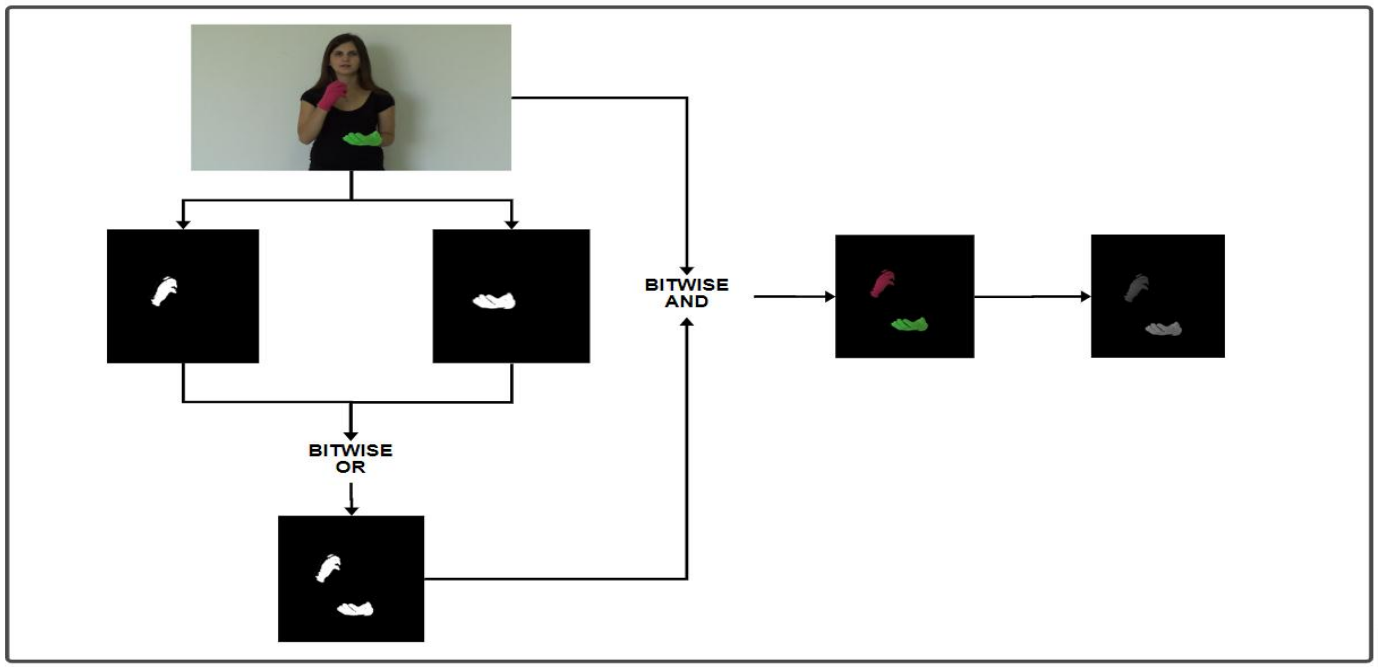

Fig-3 Extraction of hands using colour segmentation

The Frame 'F' from the video is taken and Both colours from the hands will get extracted separately, say H1 and H2. Both H1 and H2 will be merged by doing Bitwise OR operation after which we get F2., which then will be multiplied with the original image F to get a coloured image of hands (COL_FRAME) and then be converted to grayscale image (GRAY_IMG)

F => H1
F => H2
(H1 + H2) => F2
(F2 AND F) => COL_FRAME
COL_FRAME => GRAY_IMG

CNN (Convolutional Neural Networks)

Convolutional Neural Networks are a kind of neural network used in classification processes and computer vision problems. Using a basic Neural Network on images will not be compatible because an Image consists of numerous pixels, so we convolute them with convolutional filters in CNN. Using CNN we can extract spatial features from the image. CNN uses a special technique called convolution, where convolutional filters are applied to get various features of an image. In this paper, we used it to get hand spatial features. We used the transfer learning concept to train the frames using the Inception V3 model. we have taken the ImageNet dataset trained weights for the model. We have used the Inception V3 model because it requires less computational power and also keeps the performance high. The accuracy of the trained CNN model was **95%** after testing on around 85000 images.

RNN (Recurrent Neural Network)

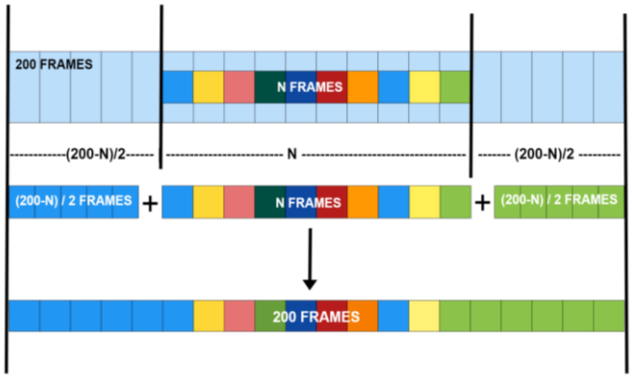

Fig-4 Padding the Frames of shorter videos

RNNs, or Recurrent Neural Networks, are a type of Neural Network that has memory and is used to evaluate time-related data, also known as Sequential Data. RNN is a type of neural network that is used in voice recognition systems like Apple's Siri and Google's voice search. Usually, Basic RNN does not do well with sequences that require memory. So, To predict the sign given a sequence of frames, we employed the Long Short-Term Memory (LSTM) Model which is a kind of RNN that has memory. The LSA64 dataset, an Argentinian Sign Language dataset,

was used to train for 63 words. We utilized RNN to assess the series of spatial features acquired from the CNN model because it is often used for sequential data. we have given obtained spatial features from CNN to RNN Model, where input is of 200 length. so we divide a video into 200 frames, a give it to the LSTM model we created.
So, in order to give an input of a specific length we need to pad videos that are smaller in length. so we used a function to pad the video that is smaller in size. we divide a video into 200 frames if the video is larger in size. we pad frames by taking first and last frames of the video and fill both ends of the video with respective frames evenly to make it of 200 length.

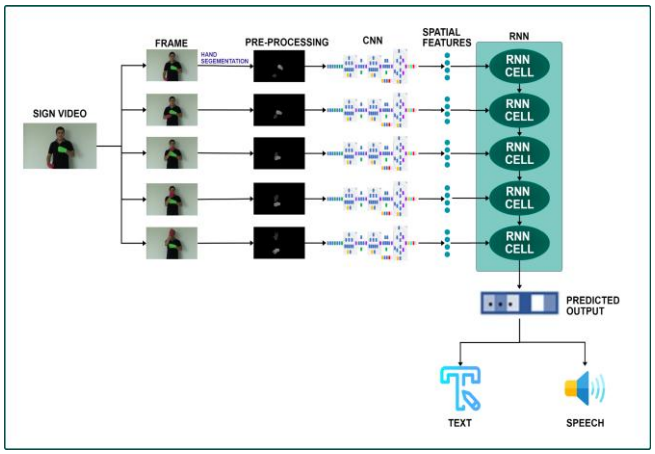

Fig-5 Architecture of the system

4 ARCHITECTURE

The architecture provides the complete process of the model from beginning to end. Figure 5 shows the entire architecture of the system. We start by taking video and divide them into frames.

CNN model we trained will be given all the frames of the video, the CNN model will output the spatial features which are in the form array. The arrays will then be padded to 200 arrays if the video is short. All the arrays will be given to the LSTM model. By taking spatial features and by doing the temporal analysis of those frames, the LSTM model will predict the sign given in the given video. The resulted output from LSTM will be converted to text and speech will be presented to the user.

5. RESULTS

Accept	10	Deaf	10	Perfume	8
Appear	10	Drawer	7	Photo	10
Argentina	9	Enemy	9	Realize	8
Away	10	Find	8	Red	10
Barbecue	10	Food	8	Rice	10
Bathe	10	Give	10	Run	10
Birthday	10	Green	10	Ship	10
Bitter	8	Help	10	Shut down	10
Born	10	Hungry	10	Skimmer	9
Breakfast	10	Last name	10	Son	8
Bright	10	Learn	10	Spaghetti	9
Buy	7	Light-blue	10	Sweet milk	10
Call	10	Man	8	Thanks	8
Candy	8	Map	10	To land	10
Catch	8	Milk	10	Trap	10
Chewing-gum	9	Mock	10	Uruguay	10
Coin	10	Music	10	Water	10
Colors	10	Name	8	Where	10
Copy	10	None	9	Women	9
Country	7	Opaque	8	Yellow	10
Dance	10	Patience	8	Yogurt	10

Fig-6 Test results

The proposed system was put to the test and found to be effective and feasible. The performance of the two models has been tested separately and combined. The CNN model which we trained on around 320000 images has been tested on 80000 images, the accuracy of the CNN was 94.91%. the accuracy of the RNN model when tested gave 97.22%. when the combined model was tested on 10 images per video which is exactly 630 videos, 588 videos out of those 630 videos have correctly predicted the sign in the video which is around 94%. Figure 6 shows the number of correctly predicted outputs per 10 videos for every sign in the dataset.

6. CONCLUSIONS

We presented a Deep-learning model, which is based on computer vision, that can recognise sign language and present it to the user with text and sound form. Rather than utilising image classification to solve sign language recognition, we approached it as a video classification problem. For this video classification problem, we used a combination of CNN and RNN models which is very efficient with an accuracy of around 94%. This can be expanded to other areas of gesture recognition as it is

providing higher performance. This can not only recognise sign language, but also other gestures which can be used in other activities.

We can extend the system to recognise continuous sign language where multiple signs can be detected using single video. for that, a semantic=c boundary detection model has to developed to separate the signs in the video and predict them separately. since the grammar is very much different from English, we can also use Natural Language Processing to convert Sign language to English with its grammar.

ACKNOWLEDGEMENT

We express our sincere gratitude to our guide, Associate Professor **Dr. Narayanamoorthy M, Dept of IoT, Vellore Institute of Technology, Vellore** for suggestions and support during every stage of this work.

REFERENCES

[1]Kehuang Li, Zhengyu Zhou, Chin-Hui Lee; Sign Transition Modeling and a Scalable Solution to Continuous Sign Language Recognition for Real-World Applications

[2]Multi-Information Spatial-Temporal LSTM Fusion Continuous Sign Language Neural Machine Translation

[3]Chengcheng Wei; Jian Zhao; Wengang Zhou; Houqiang Li, Semantic Boundary Detection With Reinforcement Learning for Continuous Sign Language Recognition

[4]D. Avola, M. Bernardi, L. Cinque, G. L. Foresti, and C. Massaroni, ‘‘Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures.’’

[5]O. Koller, H. Ney, and R. Bowden, “Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled,”

[6]N. C. Camgöz, S. Hadfield, O. Koller, and R. Bowden, “Subunets: Endto-end hand shape and continuous sign language recognition,”

[7]A Chinese sign language recognition system based on SOFM/SRN/HMM

[8]Subunit sign modelling framework for continuous sign language recognition

[10] R. Cui, H. Liu, and C. Zhang, “Recurrent convolutional neural networks for continuous sign language recognition by staged optimization,”

[11] O. Koller, S. Zargaran, and H. Ney, “Re-Sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs,”

[12] O. Koller, C. Camgoz, H. Ney, and R. Bowden, “Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos,”

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,”

[16] R. Cui, H. Liu, and C. Zhang, “A deep neural framework for continuous sign language recognition by iterative training,”

[17] T. Starner, J. Weaver, and A. Pentland, “Real-time American sign language recognition using desk and wearable computer based video,”

[18] N. Habili, C. C. Lim, and A. Moini, “Segmentation of the face andhands in sign language video sequences using color and motion cues,”

[19] L.-C. Wang, R. Wang, D.-H. Kong, and B.-C. Yin, “Similarity assessment model for chinese sign language videos,”

[21] S. Wang, D. Guo, W.-G. Zhou, Z.-J. Zha, and M. Wang, “Connectionist temporal fusion for sign language translation,”

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,”

[23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,”

[24] C. Szegedy et al., “Going deeper with convolutions,”

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,”

[26] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, “Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video,”

[27] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,”

[28] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,”

[29] J. Li, S. Zhang, and T. Huang, "Multi-scale 3D convolution network for video based person re-identification,"

[30] S. Chen, J. Chen, Q. Jin, and A. Hauptmann, "Video captioning with guidance of multimodal latent topics,"

[31] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3D convolutional neural networks,"

[32] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks,"

[33] Z. Liu, X. Chai, Z. Liu, and X. Chen, " ous gesture recognition with hand-oriented spatiotemporal feature,"

[34] D. Guo, W. Zhou, A. Li, H. Li, and M. Wang, "Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation,"

[34] D. Guo, W. Zhou, A. Li, H. Li, and M. Wang, "Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation,"

[35] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks,"

[36] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation,"

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory,"

[38] K. Cho et al., "Learning phrase representations using RNN encoderdecoder for statistical machine translation,"

[39] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical lstm for sign language translation,"

[40] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation,"