ISSN 2395-1621

SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

Manoj Lad, Kalu Mane, Priyad Khobragade, Tushar Bankar

manojlad111@gmail.com
kalumane99@gmail.com
priyadkhobragade358@gmail.com
tusharbankar.tb@gmail.com

Department of Computer Engineering, Bachelors of Engineering, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune University, Pune, India-411001.

ABSTRACT

Sign language is a communication way for deaf community. Due to the complexity the sign language is bit hard to understand. Many peoples have been researching on sign language since years back. Due to the complexity of sign language it's been making the system harder to recognize all the words accurately So there are a lot methods for sign language and to find out which one is best is the critical task. Because of the similarity between Sign language recognition and Action Recognition both are almost impossible to distinguish, so we are trying to implement one of best models in Action Recognition which is called as i3d inception model to clear all these difficulties and complexities

Keywords: Deep Learning, Convolutional Neural Networks, Sign Language.

ARTICLE INFO

Article History

Received: 8th March 2020

Received in revised form :

8th March 2020

Accepted: 10th March 2020

Published online :

11th March 2020

I. INTRODUCTION

Deaf is a descriptor word use to resemble hard of hearing or unable to hear. Difficulty in hearing tends to do the limited communication with other peoples which causes the less interaction with society. Thus, there is a strategy called Sign Language which is an excellent method to deaf people which gives them opportunity to speak with others. With this examination, we trust we can manufacture a superior correspondence with hard of hearing people groups. For instance, each nation has its own Sign Language and standard. That is the reason we need to attempt to execute the i3d model into Sign Language and break down the outcome. There is a method called Sign Language which is a very good way to make deaf and muted people to be able communicate with others. But the problem is not all peoples can understand the sign language. By using sign language one can also be not able to communicate well. To derive the simple solution and simpler way to understand sign language one of the top-tier model i3d inception is used.

So, to get rid of these difficulties there is a method called Sign Language which is a very good way to make them to be able communicate with others. But the problem is not all peoples can understand or willing to learn about Sign Language. Sign Language is very complex and consists of many hand movements. By using the Convolutional and deep neural networks it becomes more complex. Sign Language is very complex and consists of different neural networks. Even using different dataset of the same method no one can make any big difference. That's why we want to try to implement the i3d inception model into Sign Language and analyze the result using CNN.

II. LITERATURE SURVEY

Researchers are trying their best to solve the problem of this topic. There are so many researches had been done work on it so far. But there still need some improvisations. Some researchers done a very great job in sign language as some of them used ReLu. By using the JADM and integrating it with mocap videos sign language gives an optimistic result for future enhancements. With the use of Deep and Convolutional Neural Network, the sign language gives a very efficient outcomes as mentioned below:

In [1], researchers proposed 3D SignNet CNN is inspired by the VGG architecture which is a very deep and complex CNN model that demonstrated classification and localization accuracy in the ImageNet which resembles Large Scale Visual Recognition by which no supplicate data can be found. VGG is a deep Convolutional Neural Network with weighted convolutional layers having some window sizes. In which all 3D sign was represented using

upper joints of human body. The 3D template includes 57 markers:18 each for the left and right hands, 2 for the shoulders, 1 for chest, 2 for arms, 12 for the face and 4 for the head. These were given by all the datasets in a fixed pattern, starting with head before moving on to the face, chest, right fingers, left shoulder, left hand, right shoulder, right hand, and left fingers. They made a 3D SL dataset consisting of 200 Indian sign language signs, which they captured from 10 different signers/users by using a 9-cam 3D mocap system. Each signer repeated each sign at least 10 times, with varying in hand speed and trajectory, angles, movements including changes in part of the face or head that was being focused on. This gives a total of 20,000 3D sign videos, divided into 200 classes with 100 signs per class. For each class, 50 videos were chosen at random for training and remaining 50 for testing.

Research [1] proposed the use of JADMs for representing the spatio-temporal information in 3D mocap videos. This models the information by using the distances between them and angles between joint pairs. Furtherly, they proposed CNN architecture for classifying the images. They also evaluated the proposed method by comparing their performance with state-of-the-art baseline methods.

In [2], they proposed spatiotemporal two-stream network in deep learning which divides the video features into time and spatial flow to the human visual perception. Its feasibility and efficiency have been verified on multiple behavior recognition standard data sets. Combining the two-stream network and multimodal data input, if the effect of the feature extraction can be more precise and accurate, the performance index of the recognition system will be greatly improved as there is no need of any modifications. As the degree of intelligence of the system will be further strengthened. The flow is the instantaneous velocity of moving objects on the imaging plane. This motion is reflected in the movement of the pixels. It uses the changes of each pixel in the frames sequence to find relation between the two consecutive images. Optical flow method evaluates motion features of object between sequential frames. This feature requires continuous video input, and the huge calculation amount makes the speed of the whole network is reduced. Therefore, continuous sampling is difficult to cover the whole video, which restricts the development of the optical flow network.

In [3], Researchers initiated to create the dataset with five different topics performing 200 signs in 5 different angles under variety of backgrounds. Each sign occupied for 60 frames or images in a video. CNN training is performed with 3 different sample sizes which consisting of multiple sets of subjects and viewing angles. The remaining 2 samples are used for testing the trained CNN. Different Convolutional Neural architectures were designed and tested with sign language data to obtain better accuracy in identification. The dataset is having 200 ISL commonly used words performed by 5 native ISL users in 5 angles differing in their viewing at a rate of 30fps [3]. Training is initiated with three different batch sizes. In Batch-I of training only one set, in which 200 signs performed by single user in 5 different viewing angles for 2 seconds at 30fps, total sign images are 200□1□5□□2□30 □□60000. Batch-II of training is done using 2 sets which consists a total of 200□□2□5□□2□30 □120000 sign images. In Batch-III of training 3 sets of sign images were used and so on. CNN's are tested with two discrete video sets having different users and viewing angles with varying backgrounds. The robustness testing is performed in two cases. In case-1 of testing same dataset i.e. already trained dataset is used and in case-2 of testing different dataset is used. Their model consists of 6 layers including input and output in which input layer needs 2 inputs. Thus, the shape would be 2x2x64x64x32 (Grayscale and Depth x Hand and upper body x H x W x F). His proposed model is not 3D Convolutional Neural Network, so all the kernels are 2D. All the neurons are Retrificted Linea Units (ReLU). The proposed model consists of 3 layers of Convolutional Neural Network and continue by classical Artificial Neural Network or fully connected layer. With this proposed model he got 91.70% accuracy and 8.30% error rate.

In [4], Researcher brings an innovation in Action Recognition in which she did a research in which she collected some ASL videos from a website. In these collected ASL videos a person is signing some word of American Sign Language. Extraction of Frame is done from each collected sign language video. The frames of each collected ASL video are stored in a folder and the folder is named with the word which is signed by the person in the collected video. By following this procedure, many folders consisting of frames extracted from collected ASL videos are created and they are named in the same way. All these folders are stored in the dictionary folder called Collected-Videoframes. Then, they have given a video as input to the system. Frames of the inputted ASL video are extracted and these frames are stored in a folder called SignwritingInput folder. This procedure is carried out for every ASL video which is given as input.

In [5], Some authors put a research based on the images are re-sized into a size of 120x120 by maintaining the aspect ratio. They are passed through a number of convolutional layers, ReLu and max-pooling layers are reduced to 90 feature maps of size 7x7. Then they are passed on two fully-connected layers with 1024 and 144 neurons respectively. Finally, the last layer uses soft max activation for classification of gestures. Using datasets from [16], they used transfer learning method to trained model with RGB images. The weights from the convolutional layers were retained and the weights from the dense layer were re-initialized to random numbers. The results for the same is explained in the next section. Effective real time background subtraction was done using depth perception techniques. Computer vision techniques were used to gain one-to-one mapping between the depth and the RGB pixels.

In [6], A research has been done presenting CNN a large vocabulary for Sign Language Recognition. This model

benefits from the spatiotemporal feature learning capability of 3D-CNN. Instead of feeding video directly into network, they do viewpoint selection to integrate attention mechanism into model. It works as an attention seeker in the system. The spatial attention focuses on the most relevant objects and ignores background of the irrelevant parts. After feature extraction, they employed temporary attention-pooling to combine clips, where the clips are highlighted. The resulting vector can be viewed as video feature representation. They concatenate video and trajectory feature to train classifier via soft max layer. After concatenation, they got an 8192-dimensional vector as clip feature. This work focuses on isolated SLR. They will give attention to continuous SLR in future work, which translates a sign video into a semantic and continuous sentence. There are some work applying RNN for video caption that is a similar application to continuous SLR. RNN based methods provide potential to solve continuous SLR without temporal segmentation.

In [7], Researchers presented Sign Quiz, a web-based application for learning sign language with the help of Deep Neural Networks (DNN). SignQuiz application can easily be used by deaf people which help them to communicate with others. Usability, availability, low cost of operation are the factors that makes SignQuiz a useful and beneficial application for learning finger-spelled signs. With proper training this application can easily include more signs in it as much the user wants. Usability can give good results if user can select alphabet range of his own choice for learning. This will also help to understand easy or difficult signs based on the captured data. Rather than setting sign accuracy threshold, it can be set for each sign for better working. More detailed study should be done to set this. To make the application capture the sign made by the user without any external help, application is designed such as it will wait for few seconds once user clicks on the capture button. Rather than putting the delay, it shows a timer or it automatically understands that user has shown the sign and capturing it will be useful.

In [8], They initiated to design a unique spatiotemporal feature map characterization for three-dimensional (3-D) sign (or action) data. Current maps characterize geometric things, such as angles, joint distances and both, which could not accurately or more precisely modelled the relative joint variations in a 3D sign location data. Therefore, they proposed a new color-coded feature map called joint angular velocity maps to join the 3D motions. Instead of using traditional convolutional neural networks (CNNs), they propose to develop a new ResNet architecture called connived feature of ResNet, which has a CNN layer in it which densely connects standard ResNet architecture. Current 3D data models, such as JDMs and JADMs, could not characterize small motions in joints precisely. In this work, they design a novel Joint Angular Velocity maps (JAVM), which represent joint motions relatively in consecutive frames.

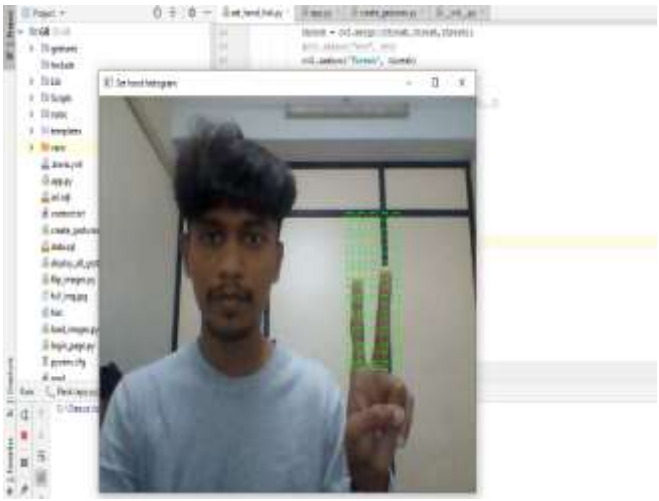
In our opinion [2],[3],[4],[5] and [6] did really well and decent to be implemented into Sign Language. As we mentioned before, Action Recognition has a similar characteristic with Sign Language.

III. MATERIALS AND METHODS

Researchers are trying their best to solve the problem of this topic. There are so many researches had been done work on it so far. But there still need some improvisations. Some researchers done a very great job in sign language as some of them used ReLu. By using the JADM and integrating it with mocap videos sign language gives an optimistic result for future enhancements.

A. Data Gathering

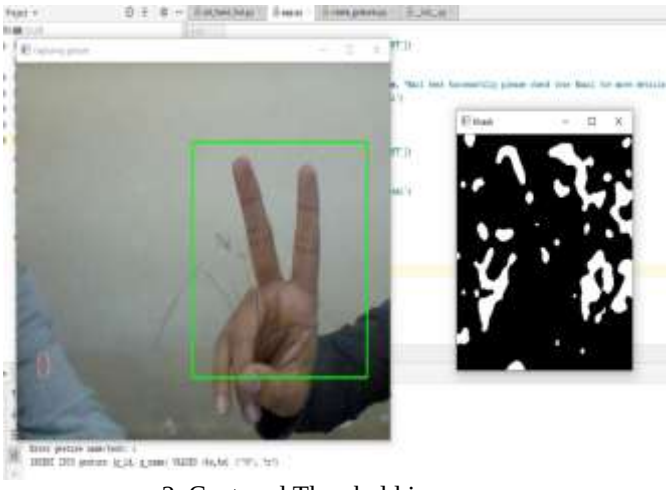
Data gathering is an important part of this research. Different datasets are affecting to the result significantly. Because of it, we used a webcam to capture images by performing the gestures accordingly.



1. Capturing gestures using Webcam

B. Data Processing

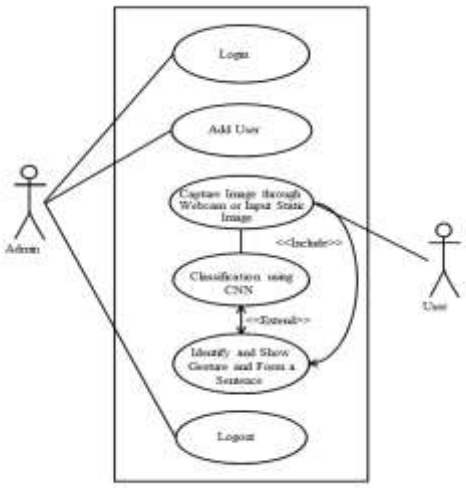
After successful capturing the images the threshold image and the original raw image will be displayed. We assign a square matrix in which the hand gesture will be captured and further processing can be done. It will check for the grayscale image by which each pixel can be identified accurately rather than RGB colors. We add blank screen which is RGB (0,0,0) for every frame that has less than 126 frames.



2. Captured Threshold image

After the convert process, we resized the frames into 224x224 height and width with bilinear interpolation. Total 1200 images will be captured and in accordance to it the train and test data will be determined. Then, pixel values were being normalized into a scale of -1 to 1. So, the numpy shape would be (1, 126, 224, 224,3). The numpy were being store into an HDF5 file using h5py library.

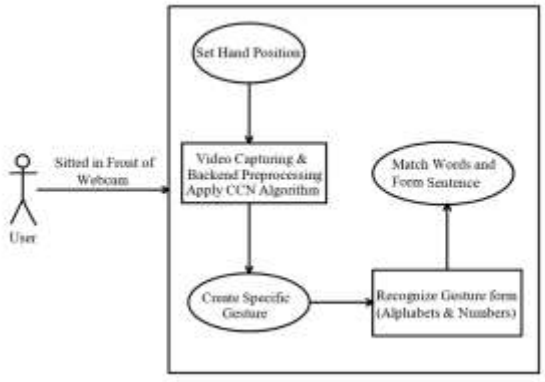
For the CNN architecture we used, it's called i3d inception or inflated inception. Because this model is based on the inception v1 model. Inception v1 was being modified into CNN. i3d inception consist of 67 convolutional layers including input and output. There are 9 inception modules. The detail of inception module is shown in Fig 2. For the training, we distribute the dataset into 7:3 ratio.



3. System Process Flow

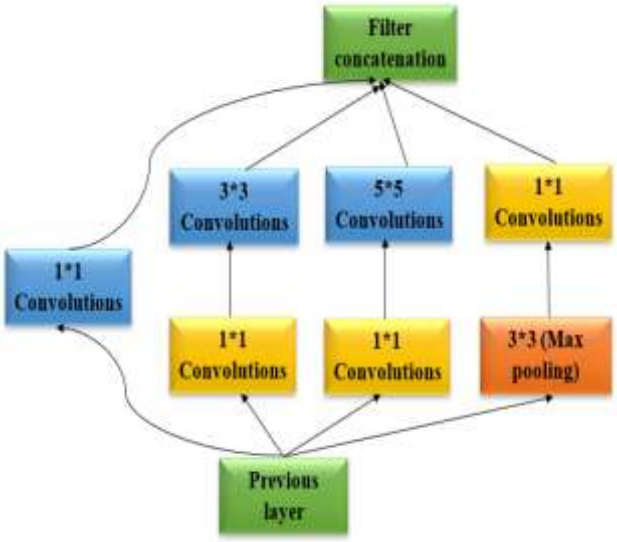
C. Convolutional Neural Network

For the CNN architecture we used, it's called i3d inception or inflated inception. Because this model is based on the inception v1 model. Inception v1 was being modified into CNN. i3d inception consist of 67 convolutional layers including input and output. There are 9 inception modules. The detail of inception module is shown in Fig 4. For the training, we distribute the dataset into 7:3 ratio.



3. System Architecture

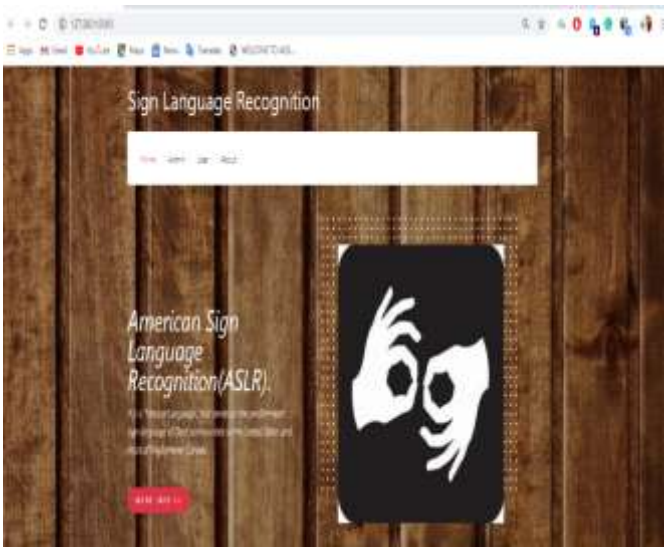
The system converts the Gestures video into simple words in English as well as make a sentence of that each word in English. The CNN process used in video processing module gives the matched results. Based on the right match, the Sign Writing Image File is retrieved and stored in a folder. This folder served as the input to Natural Language Generation Module.



4. Details of inception model

D. Webpage and security login

For creating gestures, it requires admin login for recording the hand gestures by which user will be able to understand the correct word by doing hand movements and more accurate result is interpreted.



5. Sign language webpage

6. Login page

7. Total 1200 captured images

IV. RESULTS

The result is good enough as it well captured all images and be able to display the threshold image. In this paper, we developed system for muted persons by which they can interact with normal persons. so, the American Sign Language System (ASL) introduce to manage the problem, the system applying neural networks classification approach for identifying the hand gestures like alphabets or numbers so the system can predict accurate result of hand sign and make the words as in the form of sentences.

Sr. No	Algorithm Process	Accuracy Results
1	Convolutional Neural Network Classification Technique	0.9248769

V. DISCUSSION

Due to the lighting conditions of the background sometimes it was unable to detect the hand gestures and unable to interpret the correct sign. By changing the background and keeping it plain and solid the detection is highly precise and can be identified more accurately.

REFERENCES

[1] E. Kiran Kumar, P.V.V. Kishore, “Training CNNs for 3D Sign Language Recognition with color texture coded joint angular displacement maps”, Department of Science and Technology, SEED division, SIEEE (2019).

[2] Shujun Zhang, Weijia Meng, Hui Li, Xuehong Cui, “Multimodal Spatiotemporal networks for sign language Recognition”, College of information science and technology, Qingdao University, China (2019).

[3] Aradhana Kar, Pinaki Sankar Chatterjee, “An Approach for minimizing the time taken by video processing for translating sign language to simple sentence in English”, Department of ECE, Odisha (2018).

[4] G. Anantha Rao, K. Syamala, P.V.V. Kishore, “Deep Convolutional neural networks for sign language Recognition”, Odisha, Biomechanics and vision computing Research Center, Department of ECE, SPACES (2018).

[5] Neel Kamat Bhagat, Vishnusai Y, Rathna G N, “Indian Sign Language Gesture Recognition using Image processing and Deep Learning”, Indian Institute of Science, Bengaluru, IEEE (2019).

[6] Jie Huang, Wengang Zhou, Houqiang Li, “Attention based 3D CNNs for large vocabulary Sign Language Recognition”, 1051-8215 (c), IEEE (2018).

[7] Jestin Joy, Kannan Balkrishnan, Seeraj M, “Sign Quiz: A quiz-based tool for learning finger spelled signs in Indian Sign Language using ASLR, Department of Computer applications, Cochin University of science and technology, IEEE (2018).

[8] Eepuri Kiran Kumar, P.V.V. Kishore, Maddala Teja Kiran Kumar, “3D Sign Language Recognition with Angular velocity maps and Connived feature ResNet”, IEEE-2018.

[9] Suharjito, Herman Gunawan, Ariadi Nurgroho, “Sign Language Recognition using Modified Convolutional Neural Network Model”, Master of Computer Science, Bina Nusantara University, Jakarta Indonesia (2018).

[10] Kshitij bantupalli, Ying Xie, “American sign language Recognition using Deep learning and computer vision”, Department of Computer science Kennesaw State university Kennesaw, (USA) IEEE International Conference on Big data (Big data), Sept.5-7, 2018.

[11] Garima Joshi, Renu Vig, Sukhwinder Singh, “DCA-based unimodal feature-level fusion od orthogonal moments for Indian sign language dataset”, Electronics and communication Engineering Department, UIE, IET journal-2017.

[12] Ignazio Infanito, Riccardo Rizzo, “A Framework for Sign Language Sentence Recognition by commonsense Context”, Department of Computer Engineering (DINFO), Italy (2008).
[13] Noor Tubaiz, Tamer Shanableh, “Glove based continuous Arabic Sign Language Recognition in User-Dependent Mode”, Department of Computer Engineering, American university of Sharjah, UAE-2015.
[14] M.F. Tolba, A.S. Elons, “Recent Developments in Sign language Recognition System”, Ain Shams University Cairo, Egypt (2013).
[15] Iurii Krak, Iurii Kryvonos, “Interactive systems for sign language Learning”, V.M. Glushkov Institute of Cybernetics Kyiv, Ukraine (2013).
[16] Umang Patel, Aarti G. Ambekar, “Moment based Sign language recognition for Indian languages”, DJSCE, Vile Parle Mumbai, India (2017).
[17] Md Azher Uddin Shayhan Ameen Chowdhury, “Hand Sign Language Recognition for Bangla Alphabet using Support Vector Machine”, Computer Science Engineering, Kyung Hee University South Korea (2016).
[18] Yangho Ji, Sunmok Kim, and Ki-Baek Lee, “Sign Language Learning System with Image Sampling and Convolutional Neural Network”, Kwangwoon University Seoul, South Korea, IEEE International Conference on Robotic Computing (2017).
[19] Kusumika Krori Dutta Sunny Arokia Swamy Bellary, “Machine Learning Techniques for Indian Sign Language Recognition”, International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC) [2017].
[20] Harsh Vardhan Verma, Eshan Aggarwal, Satish Candra, “Gesture Recognition Using Kinect for Sign Language Translation”, American International University –IEEE (2018).
[21] Huang, J., Zhou, W., Li H., “Sign Language Recognition using 3D Convolutional neural networks”, IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6), Turin: IEEE (2015).
[22] Dengsheng Zhang, Guojun Lu “Content-Based Shape Retrieval Using Different Shape Descriptors: A Comparative Study”, ICME, IEEE International Conference on Multimedia and Expo (ICME01)- (2001).
[23] J Kramer and L Leife, “The Talking Glove: A Speaking Aid for Nonlocal Deaf and Blind Individuals”, Proc. Of the RESNA 12th annual Conf. (2003).
[24] K. R. Linstrom and A.J. Boye, “A neural network prediction model for a psychiatric application”, International Conference on Computational Intelligence and Multimedia Applications, Washington, (2005).
[25] M. D. Skowronski ad J.G. Harris, “Automatic speech recognition using a predictive echo state network classifier”, Neural Networks, Volume 20, Issue 3, April-(2007).
[26] Peter Wray Vamplew, “Recognition of Sign Language Using Neural Networks”, Flinders University of South Australia-(2000).

