Sign Language to Text and Speech Translation in Real Time Using Convolutional Neural Network

Ankit Ojha
Dept. of ISE
JSSATE
Bangalore, India

Ayush Pandey
Dept. of ISE
JSSATE
Bangalore, India

Shubham Maurya
Dept. of ISE
JSSATE
Bangalore, India

Abhishek Thakur
Dept. of ISE
JSSATE
Bangalore, India

Dr. Dayananda P
Dept. of ISE
JSSATE
Bangalore, India

***Abstract*—Creating a desktop application that uses a computer's webcam to capture a person signing gestures for American sign language (ASL), and translate it into corresponding text and speech in real time. The translated sign language gesture will be acquired in text which is farther converted into audio. In this manner we are implementing a finger spelling sign language translator. To enable the detection of gestures, we are making use of a Convolutional neural network (CNN). A CNN is highly efficient in tackling computer vision problems and is capable of detecting the desired features with a high degree of accuracy upon sufficient training.**

***Keywords*— CNN, ASL, training**

I. INTRODUCTION

American Sign Language (ASL) is natural syntax that has the same etymological homes as being speaking languages, having completely different grammar, ASL can be express with destiny of actions of the body. In native America, people who are deaf or can't see, it's a reliable source of absurdity. There is not any formal or familiar form of sign language. Different signal languages are speculating in particular areas. For a case, British Sign Language (BSL) is an entirely different language from an ASL, and USA people who familiarise with ASL would not easily understand BSL. Some nations adopt capabilities of ASL of their sign languages. Sign language is a way of verbal exchange via human beings diminished by speech and listening to loss. Around 360 million human beings globally be afflicted via unable to hearing loss out of which 328000000 are adults and 32000000 children. hearing impairment extra than 40 decibels in the better listening to ear is referred as disabling listening to loss. Thus, with growing range of people with deafness, there is moreover a rise in demand for translators. Minimizing the verbal exchange gap among listening to impaired and regular humans turns into a want to make certain effective conversation among all. Sign language translation is one of the amongst most growing line of research nowadays and its miles the maximum natural manner of communication for the humans with hearing impairments. A hand gesture recognition gadget can offer an opportunity for deaf people to talk with vocal humans without the need of an interpreter. The system is built for the automated conversion of ASL into textual content and speech. A massive set of samples has been utilized in proposed device to understand isolated phrases from the same old American sign language which may be concerned about the use of virtual camera. Considering all the sign language alphabets and terms, the database includes one thousand special gesture images. The proposed system intends to understand some very fundamental elements of signal language and to translate them to text and audio. American Sign Language is a visible language. Along with the signing, the thoughts techniques linguistic data through the vision. The form, placement, motion of hands, in addition to facial expressions, frame movements, every play essential factor in convey facts. Sign language isn't a normal language — each the entire USA. It Has its very own signal 6 language, and areas have vernaculars, like the numerous languages are spoken anywhere inside the globally speaking language, the detection rate by the ASL language as in compare to the grammatical accuracy is of 90 % percentage of institutions commonly use Indian sign language. The amazing elements of India it [ISL] has a bit difference in signing however the grammar is identical at a few stages in the U.S.A. The Deaf humans in India remember the fact that it's plenty better than one-of-a-kind sign languages on the grounds that it's far a natural method for them, they observe via the herbal interaction with the human beings around them. The stages of sign language acquisition are equal as spoken languages, the toddlers begin with the aid of rambling with their hands. Since India doesn't have many Institutions for growing Indian sign language [other than ISLRTC which is established last year: would be future of ISL] there is lack of understanding a number of the human beings and some Institution indicates to select ASL over ISL without right knowledge.

II. LITERATURE SURVEY

Paper [1] demonstrates, a hand free demonstration of Taiwanese data language which uses the wireless system to process the data. To differentiate hand motion, they have inner sensors put into gloves to show the parameters as given by, posture, orientation, motion, defined of the hand in Taiwanese Sign Language could be recognize in no error. The hand gesture is considered by flex inner sensor and the palm size considered using the g sensor and the movement is considered using the gyroscope. Input signals would have to

be consider for testing for the sign to be legal or not periodically. As the signal which was sampled can stay longer than the pre-set time, the legal gesture sent using phone via connectivity like Bluetooth for differentiating gestures and translates it. With the proposed architecture and algorithm, the accuracy for gesture recognition is quite satisfactory. As demonstrated the result get the accuracy of 94% with the concurrent architecture.

Having a real-time sign language detector increases the efficiency of the community to able to in contact with people having disabilities like hearing and deaf society. Authors [2] have using machine learning algorithms presented the idea of a translation with skin colour tone to detect the ASL. They have made a skin colour segmentation that automatically depicts the colour and give the tune to it for further detection. They have used YCbCr spacing of colour as it's vastly used in video template code and gives the efficient results of colour tone for human skin. Further they have taken the CbCr plane to distribute the skin tone colour. People from Different ethnicity have their tones different which is crafted in a model. Deep Learning methods and several machine learning algorithms are used to demonstrate translator to translate between parties.

Authors of paper [3] applied a method of using a synthetic named animation making approach they have converted Malayalam language to Indian sign language. The intermediate representation in this method for sign language is being used by HamNoSys. In this method the application accepts some sets of words, say, either one or more and forms It in animated portion. There's an interactive system which further converts the portion of words in to HamNoSys designed structure. It's application pars everything that has been designed as it used by Kerala government to teach about sign language and subtle awareness.

Having to communicate between deaf people and normal public has become a difficult task now days and to implement a such as the society lacks a good translator for it and having an app for it in our mobile phones is like having a dream at day. Authors [4] have proposed something great for the deaf community or hearing aid community by providing an app for the communication. But making an app for it is no simple task at it requires lot of efforts like memory utilization and a perfectly fined design to implement a such. What their application does is that they take a picture of a sign gesture and later converts is to a meaningful word. At first, they have compared the gesture using histogram that has been related to the sample test and moreover samples that are obliged to BRIEF to basically reduce the weight on the CPU and its time. They have explained a process on which on their app, it's very easy to add up a gesture and store it in their database for further and expand detection set. So lastly, they came strong with having an app as a translator instead of several applications that are being used lately by users.

This paper [5] is completely based on a Spanish speaking language which converts the basic words into the Spanish language which is good for Spanish deaf people as it will provide them a stance to understand the sign language at pace as it'll be converted in a Spanish language rather than in English which is popularly used as ASL. The device or an application that they have made for this comprises of many terms such as visible interface which is used by the deaf person to specify the sequence of sign data, a translator, which simply converts those series in Spanish based language in a formed series, and convertor to speech, which basically converts those entire bits into a meaningful sentence in Spanish, of course. They mainly focus on the interface which is designed visually for the impaired which has shown many ways to writing the sign language in a real time. There are some major techniques they have used in to translate as a final system model. The mainly test data came from deaf people from cities like Madrid, Toledo and that's it as a starting data which included measurements as an important form of data information. It's been the first designed Spanish corpus for a diverse research which target only to a specific domain. Containing more than 4000 sentences of Spanish language that is later being translated into LSE. With a several editions they have finally provided a fetched version of translator with domain as renewal of Identification having records and having used pressure license.

The authors [6] have built a system which works in a continuous manner in which the sign language gesture series is provided to make a automate training set and providing the spots sign from the set from training. They have proposed a system with instance learning as density matrix algorithm that supervises the sentence and figures out the compound sign gesture related to it with a supervision of noisy texts. The set at first that they had used to show the continuous data stream of words is further taken as a training set for recognizing the gesture posture. They have experimented this set on a confined set of automated data that is used for training of them, identification for them and detection stored a subtle sign data to them. It has been stored around thirty sign language data that was extracted from the designed proposal. The Mexican Sign Language (LSM) is a language of the deaf Mexican network, which consists of a series of gestural symptoms and signs articulated thru palms and observed with facial expressions. Further the authors [7] explained, the lack of automated structure to translate symptoms from LSM makes integration of listening to-impaired human beings to society extra difficult. This painting affords a totally new technique for LSM alphanumerical signs popularity based on 3D Haar-like features extracted from depth pictures captured by means of the Microsoft Kinect sensor. Features are processed with a boosting set of policies. To observe normal performance of our technique, we Identified a tough and speedy sign from letters and numbers, and in comparison, the results with the use of traditional 2D Haar-like capabilities. Our gadget is capable of recognize static LSM signs and signs with a higher accuracy percentage than the one obtained with extensively used 2D features.

As It has been preferred for a society that having a sign language for hearing impaired and deaf people to communicate. As coming of several technologies in the

past, it's been kind of easy to have a translator which converts the sign language to the appropriate sentence and quite popular too. As they have shown this in their paper [8], it's firstly based on an Arabic sign language which automates the process of being translated on to give a subtle way of communication and further they have shown that the scope of their project apars the usage and defined set of measurements. The application directly converts the Arabic sign language into a meaningful sentence by applying an automated Machine learning algorithm as they concluded.

Authors of the paper [9][18] have presented multiple experiments to design a statistical model for deaf people for the conversion to sign language from the speech set. They have further made the system that automates the speech recognition by ASR by the help of animated demonstration and translation statistical module for multiple sets of signs. As they went ahead, they used the following approaches for the translation process, i.e., state transducer and phrase defined system. As of evaluation certain figures type have been followed: WER, BLEU after that comes the NIST. This paper demonstrates the process that translates the speech by automation recognizer having all three mentioned configurations. The paper came up with the result with finite type state transducer having the word error rate among the range of 28.21% and 29.27% for the output of ASR.

The following paper [10] depict the usage of new and standardized model of communication system in which it basically targets the deaf people as that they have further explained. The system comprised of the following two scenarios like at first, Spanish speeches to Spanish sign translation that makes the usage of a translator to break up the words and have them converted into the stream set of signs which in along with each other makes a sentence and also depicts in an avatar form. Now the second scenario, the sign language that from a speech was generated then send to the generator, what this generator does that it makes out those signs into the Spanish word which will make sense which basically converts the sign language into the Spanish based spoken language words. Now whatever the data previously was generated is comprises of words, now the last step is to convert the words into a sentence that is word to speech conversion. This paper consists of real example in city of Spain, Toledo that involved government personal and the deaf people. Further the paper describes the possible outcome and the scope of it in the future.

In the following article [11] it depicts the communication between deaf people and the other parts of society. It further depicts the movement of body, hands, fingers and other factors such as emotions on the face. The motion capture has been used to transmit the movements and for the sign language depiction. The dactyl-based modelling alphabet tech is developed using 3-d model. Having the efficiency on the Ukrainian sign language the realization is developed. This is based for the universal character and designed for further model.

As times are improving and developing sign language emerges as one the best medium to communicate with the deaf community by having the supportive system to be in a help to society. The project they have demonstrated [12] here are trying to make the communication easy by having the sign mainly having dynamic and static in ISL are being converted into the speech. A placed sensor of glove with sensor of flex help to design the orientation of hand and following actions. Using wireless transmission which converts it to the further bits of speech as the output. In this project they studied about LSTM networks which for long time formed dependencies. The result of this projects leads to a success rate of 98% accuracy which could able to identify the 26 gestures.

III. DESCRIPTION OF THE METHOD

Real time signal language to textual content and speech translation, specifically: 1. Reading man or woman signal gestures 2. Training the system learning model for image to textual content translation three. Forming words 4. Forming sentences 5. Forming the entire content 6. Obtaining audio output.

A. Flow Diagram

The flow chart explains the steps occurring to accomplish

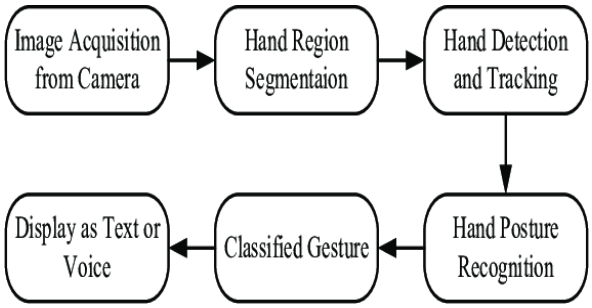

Fig 3.1 Flow Chart of the Project

the objectives of the project. These steps have been explained in a greater detail below:

1. Image Acquisition

The gestures are captured through the web camera. This OpenCV video stream is used to capture the entire signing duration. The frames are extracted from the stream and are processed as grayscale images with the dimension of 50*50. This dimension is consistent throughout the project as the entire dataset is sized exactly the same.

2. Hand Region Segmentation & Hand Detection and Tracking The captured images are scanned for hand gestures. This is a part of preprocessing before the image is fed to the model to obtain the prediction. The segments containing gestures are made more pronounced. This increases the chances of prediction by many folds.

3. Hand Posture Recognition

The preprocessed images are fed to the keras CNN model. The model that has already been trained generates the predicted label. All the gesture labels are assigned with a probability. The label with the highest probability is treated

to be the predicted label.

4.Display as Text & Speech

The model accumulates the recognized gesture to words. The recognized words are converted into the corresponding speech using the pyttsx3 library. The text to speech result is a simple work around but is an invaluable feature as it gives a feel of an actual verbal conversation.

B. Convolutional Neural Network for Detection

CNN are a class of neural network that are highly useful in solving computer vision problems. They found inspiration from the actual perception of vision that takes place in the visual cortex of our brain. They make use of a filter/kernel to scan through the entire pixel values of the image and make computations by setting appropriate weights to enable detection of a specific feature.

The CNN is equipped with layers like convolution layer, max pooling layer, flatten layer, dense layer, dropout layer and a fully connected neural network layer. These layers together make a very powerful tool that can identify features in an image. The starting layers detect low level features that gradually begin to detect more complex higher-level features.

C. The CNN Architecture functioning

The CNN model for this project consists of 11 layers. There are 3 convolutional layers. The first convolutional layer, which is responsible for identifying low level features like lines, accepts an image with 50*50 size in the grayscale image. 16 filters of size 2*2 are used in this layer which results in the generation of an activation map of 49*49 for each filter which means the output is equivalent to 49*49*16. A rectifier linear unit (relu) layer is also added to eliminate any negative values on the map and replace it with 0. A maxpooling layer is applied which reduces the activation to 25*25 by only considering maximum values in 2*2 regions of the map. This step increases the probability of detecting the desired feature. This is followed by a second convolutional layer. It is responsible for identifying features like angles and curves. This layer has 32 filters of size 3*3 which results in the generation of an activation map of 23*23 which means the output is equivalent to 23*23*32. A maxpooling layer further reduces the activation map to 8*8*32 by finding the maximum values in 3*3 regions of the map. A third convolutional layer is used to identify high level features like gestures and shapes. 64 filters of size 5*5 reduce the input to an output of 4*4*64. A maxpooling layer reduces the map to 1*1*64. The map is flattened to a 1d array of length 64. A dense layer expands the map to an array of 128 elements. A dropout layer drops out random map elements to reduce overfitting. In the end, a dense layer reduces the map to an array of 44 elements which represent the number of classes.

Each class has a corresponding probability of prediction allocated to it. The class with the maximum probability is displayed as the predicted gesture.

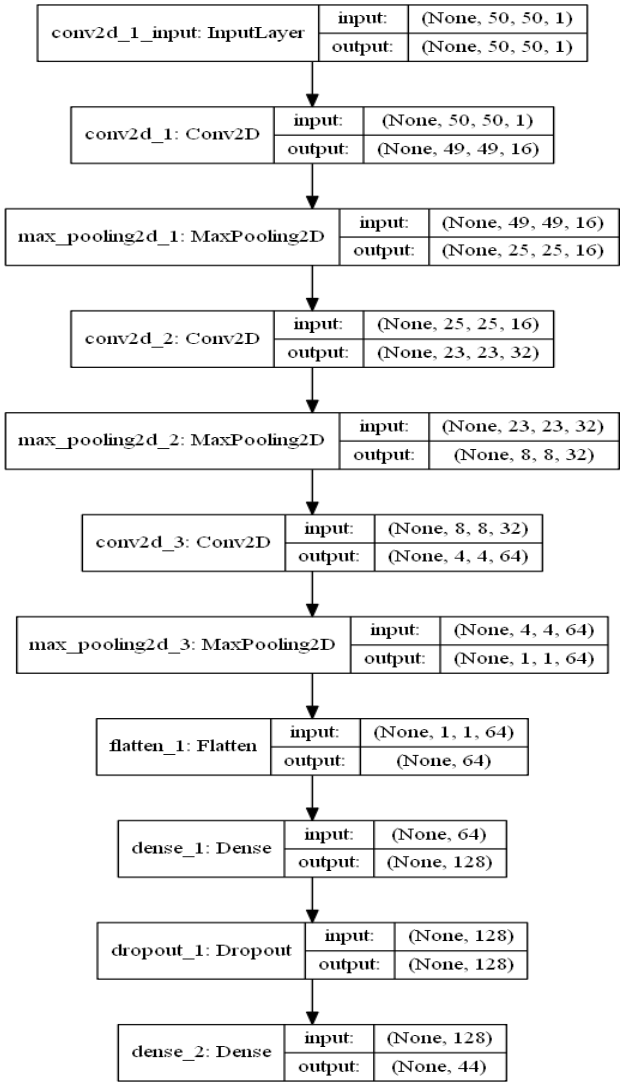

Fig 3.2 The CNN Architecture for the project

D. Recognition of Alphabets and number

To discover bounding packing containers of various objects, as we used the Gaussian historical past subtraction which used a technique to version each history pixel with the resource of a mixture of K Gaussian set distributions (k varies from 3 to 5). The possibly historical past colorations are those that stays longer are greater the static. On those fluctuating pixels, we design a square bounding field. After Obtaining all the gesture and heritage, a Convolutional NN model has designed using those photos to apart the gesture symptoms and signs from the historical beyond. These function maps explain that the CNN can understand the common unexposed structures some of the gesture indicators within training set,

and then therefore able to distinguish amongst a gesture and the past.

Fig 3.3 The Gesture Symbols for ASL Alphabets that will be in the training data

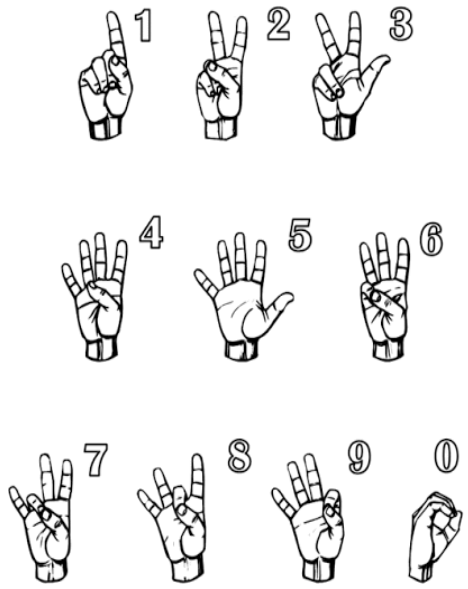

Fig 3.4 The Gesture Symbols for ASL Numbers that will be in the training data

E. Algorithm

Algorithm Real time sign language conversion to text and Start

S1: Set the hand histogram to adjust with the skin complexion and the lighting conditions.
S2: Apply data augmentation to the dataset to expand it and therefore reduce the overfitting.
S3: Split the dataset into train, test and validation data sets. S4: Train the CNN model to fit the dataset.
S5: Generate the model report which includes the accuracy, error and the confusion matrix.
S6: Execute the prediction file – this file predicts individual gestures, cumulates them into words, displays the words as text, relays the voice output.
Stop

IV. CONCLUSIONS

The project is a simple demonstration of how CNN can be used to solve computer vision problems with an extremely high degree of accuracy. A finger spelling sign language translator is obtained which has an accuracy of 95%. The project can be extended to other sign languages by building the corresponding dataset and training the CNN. Sign languages are spoken more in context rather than as finger spelling languages, thus, the project is able to solve a subset of the Sign Language translation problem. The main objective has been achieved, that is, the need for an interpreter has been eliminated. There are a few finer points that need to be considered when we are running the project. The thresh needs to be monitored so that we don't get distorted grayscales in the frames. If this issue is encountered, we need to either reset the histogram or look for places with suitable lighting conditions. We could also use gloves to eliminate the problem of varying skin complexion of the signee. In this project, we could achieve accurate prediction once we started testing using a glove. The other issue that people might face is regarding their proficiency in knowing the ASL gestures. Bad gesture postures will not yield correct prediction. This project can be enhanced in a few ways in the future, it could be built as a web or a mobile application for the users to conveniently access the project, also, the existing project only works for ASL, it can be extended to work for other native sign languages with enough dataset and training. This project implements a finger spelling translator, however, sign languages are also spoken in a contextual basis where each gesture could represent an object, verb, so, identifying this kind of a contextual signing would require a higher degree of processing and natural language processing (NLP). This is beyond the scope of this project.

V. REFERENCES

[1] L. Kau, W. Su, P. Yu and S. Wei, "A real-time portable sign language translation system," 2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS), Fort Collins, CO, 2015, pp. 1-4, doi: 10.1109/MWSCAS.2015.7282137.

[2] S. Shahriar et al., "Real-Time American Sign Language Recognition Using Skin Segmentation and Image Category Classification with Convolutional Neural Network and Deep Learning," TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018, pp. 1168-1171, doi: 10.1109/TENCON.2018.8650524.

[3] M. S. Nair, A. P. Nimitha and S. M. Idicula, "Conversion of Malayalam text to Indian sign language using synthetic animation," 2016 International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, 2016, pp. 1-4, doi: 10.1109/ICNGIS.2016.7854002.

[4] M. Mahesh, A. Jayaprakash and M. Geetha, "Sign language translator for mobile platforms," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, 2017, pp. 1176-1181, doi: 10.1109/ICACCI.2017.8126001.

[5] S. S Kumar, T. Wangyal, V. Saboo and R. Srinath, "Time Series Neural Networks for Real Time Sign Language Translation," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, 2018, pp. 243-248, doi: 10.1109/ICMLA.2018.00043.

[6] D. Kelly, J. Mc Donald and C. Markham, "Weakly Supervised Training of a Sign Language Recognition System Using Multiple Instance Learning Density Matrices," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, no. 2, pp. 526-541, April 2011, doi: 10.1109/TSMCB.2010.2065802.

[7] J. Jimenez, A. Martin, V. Uc and A. Espinosa, "Mexican Sign Language Alphanumerical Gestures Recognition using 3D Haar-like Features," in IEEE Latin America Transactions, vol. 15, no. 10, pp. 2000-2005, Oct. 2017, doi: 10.1109/TLA.2017.8071247.

[8] M. Mohandes, M. Deriche and J. Liu, "Image-Based and Sensor-Based Approaches to Arabic Sign Language Recognition," in IEEE Transactions on Human-Machine Systems, vol. 44, no. 4, pp. 551-557, Aug. 2014, doi: 10.1109/THMS.2014.2318280.

[9] R. San Segundo, B. Gallo, J. M. Lucas, R. Barra-Chicote, L. F. D'Haro and F. Fernandez, "Speech into Sign Language Statistical Translation System for Deaf People," in IEEE Latin America Transactions, vol. 7, no. 3, pp. 400- 404, July 2009, doi: 10.1109/TLA.2009.5336641.

[10] V. Lopez-Ludena, R. San-Segundo, R. Martin, D. Sanchez and A. Garcia, "Evaluating a Speech Communication System for Deaf People," in IEEE Latin America Transactions, vol. 9, no. 4, pp. 565-570, July 2011, doi: 10.1109/TLA.2011.5993744.

[11] I. Krak, I. Kryvonos and W. Wojcik, "Interactive systems for sign language learning," 2012 6th International Conference on Application of Information and Communication Technologies (AICT), Tbilisi, 2012, pp. 1-3, doi: 10.1109/ICAICT.2012.6398523.

[12] E. Abraham, A. Nayak and A. Iqbal, "Real-Time Translation of Indian Sign Language using LSTM," 2019 Global Conference for Advancement in Technology (GCAT), BANGALURU, India, 2019, pp. 1-5, doi: 10.1109/GCAT47503.2019.8978343.

[13] [13] Pigou L., Dieleman S., Kindermans PJ., Schrauwen B. (2015) Sign Language Recognition Using Convolutional Neural Networks. In: Agapito L., Bronstein M., Rother C. (eds) Computer Vision - ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science, vol 8925. Springer, Cham

[14] Bheda, Vivek and Dianna Radpour. "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language." ArXiv abs/1710.06836 (2017): n. pag.

[15] M.V, Beena. "Automatic Sign Language Finger Spelling Using Convolution Neural Network : Analysis." (2017).

[16] Tolentino, Lean Karlo S. et al. "Static Sign Language Recognition Using Deep Learning." International Journal of Machine Learning and Computing 9 (2019): 821-827.

[17] Umang Patel and Aarti G. Ambedkar, "Moment Based Sign Language Recognition for Indian Language" 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA).

[18] Bhargav Hegde, Dayananda P, Mahesh Hegde, Chetan C, "Deep Learning Technique for Detecting NSCLC", International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-3, September 2019, pp. 7841-7843