Examination of the Structure and Evolution of a Deep Learning Framework for Converting Sign Language to Spoken English

[1]R.Sagar, [2]Y.Susheela, [3]A.Sowjanya, [4]Katukojula Shivaprasad
[1,2,3]Associate Professor, [4]UG Student, [1,2,3,4]Department of Computer Science Engineering, Vaageswari College of Engineering, Karimnagar, Telangana, India

ABSTRACT
Communication between the speech- and hearing-impaired and the majority of the speaking world is hampered by the lack of popularity of sign languages among the speaking world. These languages are classified as natural languages having unique lexicons and grammars, making it challenging for the average individual to comprehend them. This work presents a hand gesture classification, monitoring, feature extraction, and categorization-based sign language recognition system. Technology helped by artificial intelligence is being utilized to reduce this communication gap and raise the standard of living for these minority. In order to facilitate the translation of sign languages into spoken languages and vice versa, a number of technologies have been suggested and created. Despite efforts in this direction, technologies for everyday use have yet to be developed and popularized. The text or speech to be signed translatable to address this issue, we propose a novel convolutional neural network (CNN) that automatically extracts discriminative spatial-temporal features from images without any prior knowledge, thereby avoiding feature design. CNN is now actively used to solvea variety of problems like detecting human activity and detection of vehicle, network intrusion detection, etc. We validate the proposed model on a real dataset collected with real-time Open CV image capture and show how it outperforms traditional approaches based on hand-crafted features. To test the proposed hybrid model, we used our own American Sign Language finger spelling dataset. This database is made up of 7000 different types of signs with each letter containing about 300 images. The inputs would be taken as still images of signs taken with the computer webcam that do not involve any motion. Furthermore, additional improvements can be done in the future to the application. It could be created as a web or smart phone application to make it easier for users to access the project. Also, while the current project only works with ASL, it might be modified to function with other native sign languages with enough data and training.
Keywords— Deep Learning; Hand Gestures; Sign Language; Speech

INTRODUCTION
Even though there are more than 20 lakh people with speech impairments and over 50 lakh people with hearing impairments in India alone, sign language has never been taught or learned by anyone, which leaves a lot of people with these impairments feeling excluded and reduces their opportunities to express their ideas and ideologies. By 2050, it's anticipated that about 2.5 billion individuals would have some degree of hearing loss, with at least 700 million of those needing hearing rehabilitation. Around 1 billion young people are at risk of preventable, permanent hearing damage due to risky listening habits. Despite technological advancements, sign language still has to be made more widely known and accessible to the general public. Engineers are working on developing a glove with sensors attached to it that can recognize hand gestures made that are made by the person wearing these gloves. Though the idea, functionality and accuracy of the device are brilliant, we need easier ways that are readily available and are simple to be used by both disabled and non-disabled to create a bridge of communication between them. A sign language is a visual language which is expressed by hand movements. Sign languages usually differ from each other based on their location. Some of them are American sign language (ASL), British sign language and Indian sign language. To form its words, ASL employs hand shape, position, movement, gestures, facial expressions, and other visual cues. People with hearing disabilities who use sign language, or "sign," can communicate rapidly and effectively with those who do the same. To go around, most deaf persons utilize a combination

of sign language, lip reading, and written communication. Many resources have been developed in the United States to assist deaf persons living regular lives. ASL is now one of the most popular languages taught on college campuses. A person familiar with American sign language may not be able to understand British sign language and vice versa.

LITERATURE SURVEY

Worldwide, different innovations have been developed to help in the translation of sign languages. Some of these technologies arebased on hardware, while others are based on software. Reference [1] exhibits systems that have been proposed and developed in which the signer wears glove-like extensions that enable gestures to be captured, identified, and translated. These gloves aid in the accurate capture of the signer's hand motions, which are then interpreted using suitable algorithms. Sensor gloves are typically made of cloth and equipped with sensors. It uses about seven sensors. Each finger and thumb have five sensors. One sensor is used to take measurements, one sensor for rotation and one sensor for hand tilt of the palm. Optical fibers are attached to the surface measuring the flexure of the fingers and toes with a glove thumb. Each sensor outputs a single digit value. A number between 0 and 4095 This value informs you about the bent of the sensor. 0 indicates that the object is fully stretched, whereas 1 indicates that it is not fully stretched. The down side of such technologies is that they can only be used once, not the signs, but the hand movements evoked by facial expressions and body language. Also, utilization of third-party devices, usage of batteries that are in close proximity to the user, motion and hand gesture recognition sensors are quite expensive.

[2] Microsoft‘s Kinect is a family of motion-sensing input devices. The devices often include RGB cameras as well as infrared projectors and detectors that are used to map depth through structured light or time of flight calculations, enabling real-time hand gesture and body skeleton identification, among other things. Microphones are also included, which can be used for speech recognition and voice control. It consists of a high-definition video camera in the that can deliver up to 1280x1024 resolution at 30frames per second. The Kinect depth sensor measures the depth of objects in the scene in front of the sensor using an infrared projector and an infrared camera. It is also designed to cancel background noise. Kinect was created as a motion controller for Xbox video game systems, and it differed from other competitors like Sony PS and Nintendo as it did not require any physical components. This sensor can read up to six skeletons at once, and it‘s small object detection is believed to be better. The Kinect sensor‘s second iteration can also see faces, track eye movements, and recognize expressions. By integrating information from depth sensors and a normal RGB camera, Kinect is able to capture the surrounding scene in 3D. This combination produces an RGB (red, green and blue) image with a resolution of 640x480, in which each pixel is assigned a color and depth information.

[3] Google released a new application called Google Gesture which helps us convert sign language to speech using an electronic wristband. Here, the signer uses a wristband in order to translate sign language into voice, the electronic wristbands detect the wearer's muscular activity while also recognizing converting the sign language to speech through and android device.

[4] Many more technologies like transformers were also discovered for sign language translation. Here, the authors present anovel end-to-end sign language translation system combining Spatial-Temporal Multi-Cue (STMC) and Transformer networks, as well as a wide range of experimental results for various Transformer setups. A tokenization method generates glosses from sign language videos first in sign language translation. The identified glosses are then translated into spoken language by a translation system.

[5] Other innovations similar to Google Gesture, have also been made. Cheng Zhang, a Cornell University professor of information sciences, along with Yin Li, a UW School of Medicine and Public Health professor of biostatistics and medical informatics, developed the software that powers FingerTrak. FingerTrak allows us track the human hand and has useful applications in sign language translation, virtual reality, mobile health, and human-robot interactions. It can detect and translate numerous positions of the human hand, including 20 finger joint positions, into 3D. The device consists of four little cameras which are used to take photos to generate a hand outline. The virtual hand is then reconstructed in 3D using a deep neural network

that stitches these silhouette images together. In this way, the entire hand pose can be captured.

PROPOSED SYSTEM

The system proposed takes real-time images using a webcam and provides an output based on the captured video. Here, we specifically deal with ASL‘s stationary alphabets from A to Z, as this is an initial stage of development, complex grammatical structures which involve motion that cannot be recognized. The system mainly follows the given steps shown in figure 1.

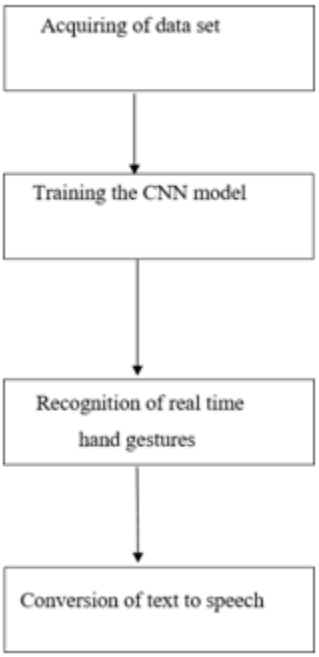


Fig. 1 Proposed System

The above figure is elaborated and explain briefly.

A. Acquiring the dataset

In acquiring or preparing a data set we create a directory that consists of subdirectories label from A to Z these directories consist of grayscale images which are acquired by capturing images using OpenCV. The data set used here consists approximately of 7000 images in JPEG format. Here, image processing is done to remove any other objects from the frame. The picture captured in ROI is firstly converted into a greyscale image using cv2.cvtColorfunction and then we use cv2.threshold to adjust the thresholdof the image which is set to 10. The contours of the hand figure are found using cv2.findContours and when contours are ascertained they are saved in different folders. The below figure shows the dataset and the subdirectory for the alphabet ‗A‘.

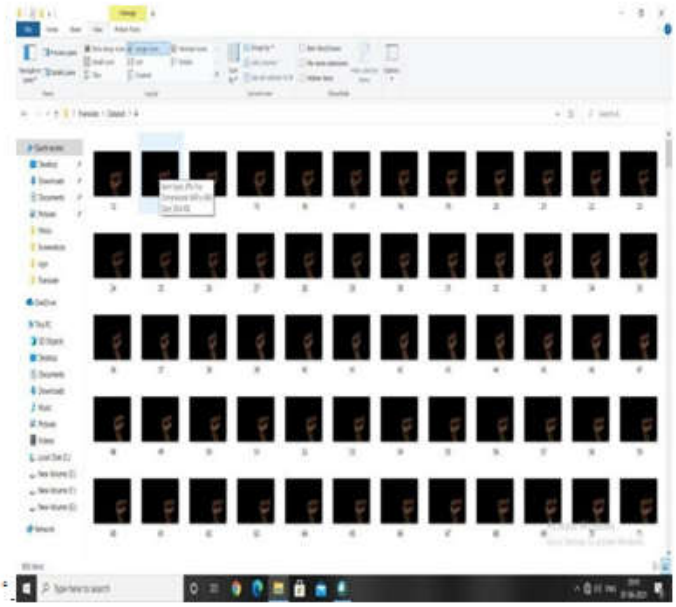


Fig. 2 Dataset for the alphabet A(sub directory)

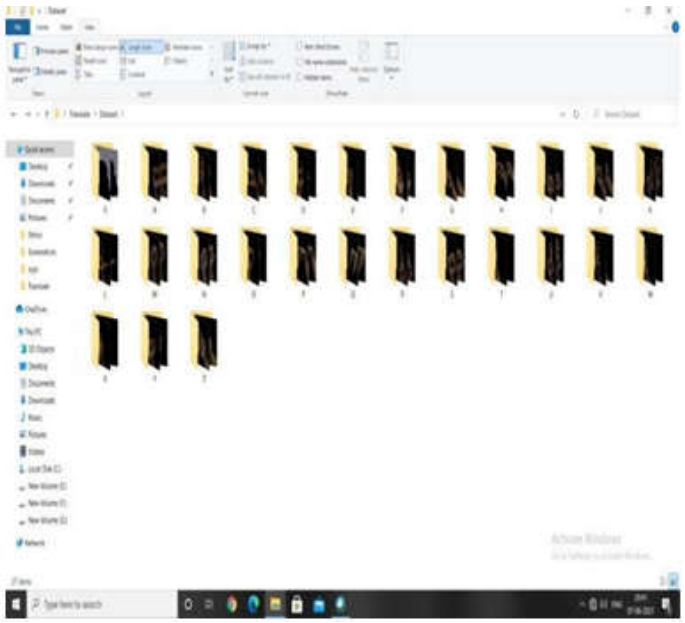
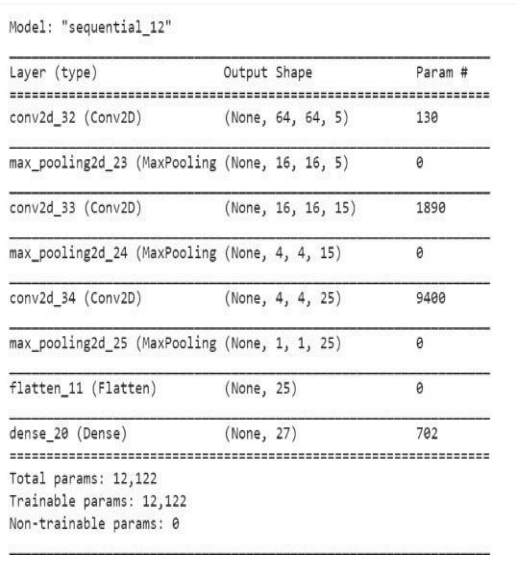


Fig. 3 Dataset Directory

B. Training the CNN model

We use Keras sequential model for recognizing hand gestures, data set is fed to the sequential model and resized to 64 x 64 which is the size that is accepted by the first convolutional layer. The testing size taken here is 0.2 which makes 20% of the data set. Training is done using the remaining 80% of the data set. We fit the model, save this model in HDF5 format so that it can be used later for real-time recognition. Accuracy and loss of the model are calculated after each epoch. The model summary is shown in figure 3.



```
Model: "sequential_12"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_32 (Conv2D)           (None, 64, 64, 5)         130
_________________________________________________________________
max_pooling2d_23 (MaxPooling (None, 16, 16, 5)         0
_________________________________________________________________
conv2d_33 (Conv2D)           (None, 16, 16, 15)        1890
_________________________________________________________________
max_pooling2d_24 (MaxPooling (None, 4, 4, 15)          0
_________________________________________________________________
conv2d_34 (Conv2D)           (None, 4, 4, 25)          9400
_________________________________________________________________
max_pooling2d_25 (MaxPooling (None, 1, 1, 25)          0
_________________________________________________________________
flatten_11 (Flatten)         (None, 25)                0
_________________________________________________________________
dense_20 (Dense)             (None, 27)                702
=================================================================
Total params: 12,122
Trainable params: 12,122
Non-trainable params: 0
_________________________________________________________________
```

Fig. 4 Model Summary

1) CNN model functioning:

Here, the CNN model considered has 8 layers of which three are convolutional. The first convolutional layer consists of 5 filters and takes the input image of a size 64 x 64 producing 130 parameters (number of filters*(width of the filter*height of the filter*number of previous layer filter +1)(5*(5*5*1) + 1 == 130) where 1 denotes the bias for each filter. A rectifier linear unit layer is also added to the before layer to remove negative values. Max pooling layer is applied to each filter(5 filters) and the data shape is decreased to 16 x 16 as the size of the pool is 4 x 4. The second convolutional layer consists of 15 filters, this layer is followed by the same relu layer and a max-pooling layer which decreases the size of the output shape to 4 x 4, the next convolutional layer has 25 filters followed by flatten and dense layers.

C. Recognition of real time gestures

We use the web camera to take real-time videos we first check if the hand exists in the frame and then draw a grey image from the frame which is re-sized and reshaped. The image captured is then converted into a threshold image which is given as an input parameter to the model. The saved model‘s predict function is used to predict the output label which is then stored in a variable and then appended when there is a change in the hand gesture.

D. Conversion from text to speech

It is a very simple task to convert from text to speech by using the gtts package available in python. We store the output predicted by the model in a variable which then is passed to the gtts function and saved in mp3 format. This saved audio is then passed to the play sound to be played. The concept of threads has been utilized to continue this process without ending the program.

RESULT ANALYSIS

A. Model analysis

This system is trained using training data. The training data comprises 80% of the entire data that is collected. This data is then divided to form validation data which is not used in training the model but is only used for validating. The graph plots the trainingand validation data that indicate over-fitting or under-fitting. These graphs are also known as learning curves.

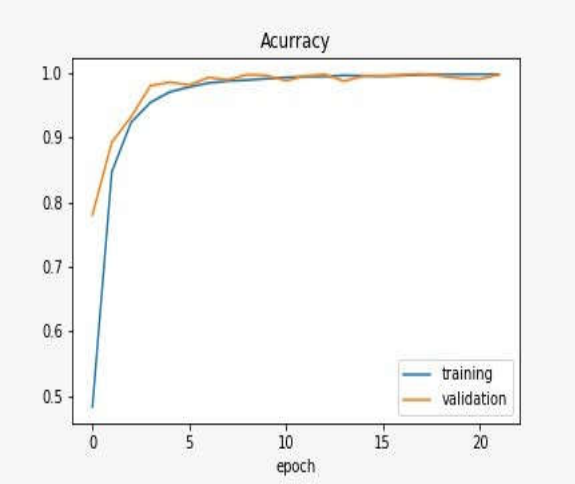


Fig. 5 Performance Learning Curves

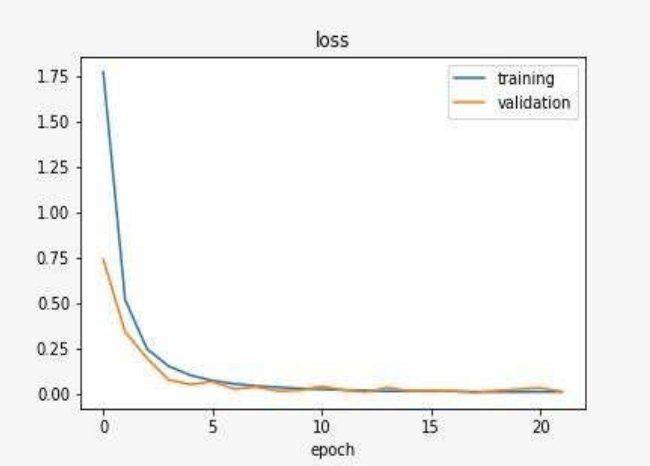


Fig. 6 Optimization Learning Curves

Here, we consider two learning curves one is optimization learning curve (fig 5) and another is performance learning curve(fig 4). The system is built with the number of epochs as 22 and the validation split is taken as 0.2, the learning curves here show a good fit learning curve.

B. Results obtained

The Model developed here when evaluated has given the following results.

```
[ ] resultmodel.evaluate(test_data,test_label)    #for check test score

152/152 [==============================] - 5s 31ms/step - loss: 0.0059 - accuracy: 0.9986
```

Fig. 7 Model Evaluation

The main purpose of this system is identifying the hand gestures and converting the textual output to speech. The identification of different hand gestures is shown below.

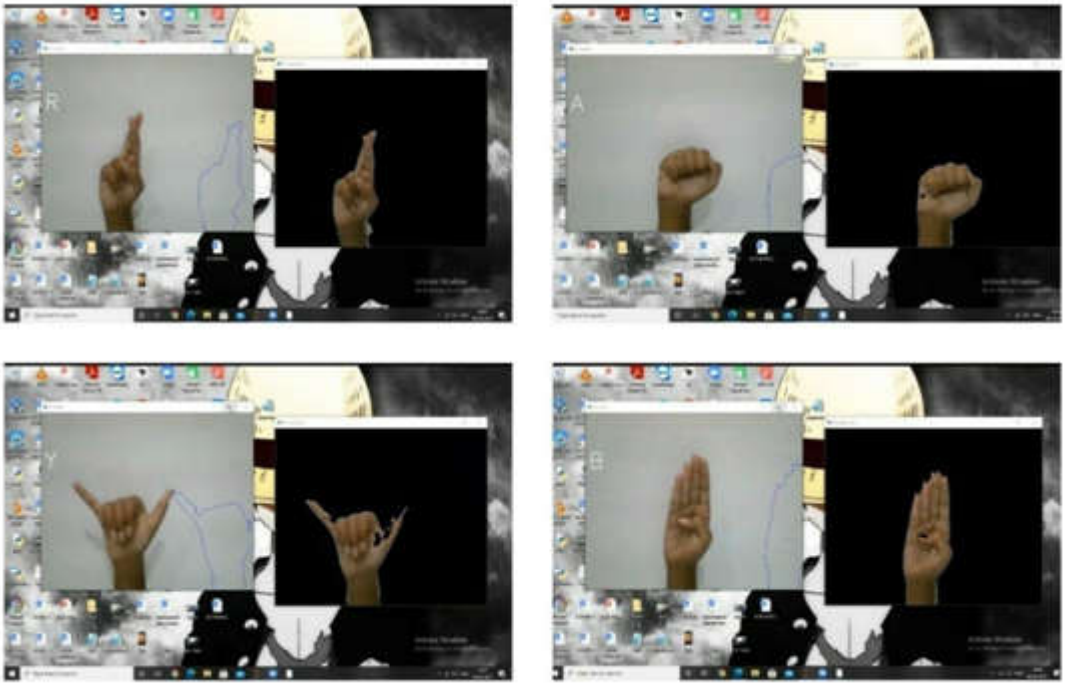


Fig. 8 Screenshots of Results obtained

CONCLUSIONS

This research is a clear demonstration of how CNN may be used to handle computer vision problems with pinpoint accuracy. The main goal was to eliminate the need for an interpreter, and that was achieved successfully. However in future, there is a scope for improvement in this project. For people to have easier access to the project, it might be built as a web or smartphone application. Because sign languages are spoken in context rather than as finger spelling languages, the project is able to address a sub-section of the Sign Language translation problem.

Furthermore, the current project only works for a small segment of the cases in the data set; however, with enough data and training, it could be expanded to work for other native sign languages. With the help of real-time photos, the system recognizes hand motions and translate the text output to speech. This project makes communicating with people having hearing impairment a

lot easier without the need to learn it or use an additional device. Sentiment analysis of the signer is another part of development. Incorporating the signer's intent or feeling alongside the gesture he or she is making would result in a more accurate system that takes into account the signer's sentiment or emotion. Because certain indicators are dependent on these elements, this could potentially aid improve translation accuracy. Including these characteristics and refining the model based on them would ideally result in an even more flexible communication in the actual world.

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), 2015, pp. 1-4, doi: 10.1109/MWSCAS.2015.7282137.
2. S. Rajaganapathy, B. Aravind, B. Keerthana, M.Sivagami, Conversation of Sign Language to Speech with Human Gestures, Procedia Computer Science,Volume 50, 2015, Pages 10-15, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.04.004.
3. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies Volume 4, Issue 2, June 2020 771 pages EISSN:2474-9567DOI:10.1145/3406789
4. Yin, Kayo and Jesse Read. ‒Attention is All You Sign: Sign Language Translation with Transformers.‖ (2020).
5. Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4, 2, Article 71 (June 2020), 24 pages.DOI:https://doi.org/10.1145/3397306
6. G. A. Rao, K. Syamala, P. V. V. Kishore and A. S. C. S. Sastry, "Deep convolutional neural networks for sign language recognition," 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES), 2018, pp. 194-197, doi: 10.1109/SPACES.2018.8316344.
7. Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9
8. Citation: Kanan C, Cottrell GW (2012) Color-to-Grayscale: Does the Method Mattler in Image Recognition? PLOS ONE 7(1): e29740.https://doi.org/10.1371/journal.pone.0029740.
9. Sreenivas, Arvind & Maheshwari, Mudit & Jain, Saiyam & Choudhary, Shalini & G, Dr. Vadivu. (2020). Indian Sign Language Communicator Using Convolutional Neural Network. International Journal of Advanced Science and Technology. 29. 11015-11031.