*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 10 Issue 5 May 2022

Converting Youtube Video to American Sign Language Translation Using Convolution Neural Network and Video Processing

Meet Modi
Hiranandani Foundation School
Hiranandani Gardens, Powai, Mumbai

Reetu Jain
Chief Mentor
On My Own Technology Pvt Ltd, Mumbai, India

Abstract:
Sign language is a form of communication used by the deaf population of the world with others or amongst the deaf population. At an estimate 5% of the world population is either deaf or suffers from hearing loss. It is often observed that the youtube videos, although the subtitles are available in English or other native languages, do not have any SL based subtitles. Therefore the comprehensive intention of the present study is to develop an American Sign Language (ASL) based subtitles for the youtube videos. The proposed method is a three phase framework that not only automates the process of youtube video and its transcript downloading but also automatically converts the text in the transcript to ASL based subtitles and mount that on the video. The proposed method is an integration of deep learning based Convolution Neural Network (CNN) and image and video processing techniques. A torch based CNN model is developed and coded in Python 3.8.5. The model showed training and testing accuracy of 99.982% and 98% respectively. The strength of a model lies in its ability to be applied in a practical problem. Therefore, the proposed integrated method is applied to extract a random video from youtube.

Keywords: Youtube video; American Sign Language; Image processing; Video processing; Convolution neural network; subtitles

1. Introduction

Sign language (SL) is a form of communication used by the deaf population of the world with others or amongst the deaf population or with the deaf population. There are various types of SL such as American Sign Language (ASL), British sign language, Indian sign language ets. Out of all the forms of SLs, ASL is the the most preferred SLs because it uses at an approximation of 6,000 gestures for common words and finger spelling recondite words or proper nouns [1]. In the United States alone around 250,000 - 500,000 people use ASL for their daily communication [2]. Although a significant percentage of the population uses SLs for communication with the deaf people, yet there is very little knowledge about it. In the youtube community, there exists very little work that converts the subtitles of a video to ASL. Therefore, the present study is adopted with the comprehensive intention to automate the process of converting the transcript of the youtube video into ASL and then append it on the downloaded video.

1.1. Motivation and Novelties

The study is divided into three phases. The first phase involves the development of a multiclass classifier to identify the English alphabets and then map them to the pictures of the alphabets in ASL. In this process, a keras and tensorflow based convolution neural network (CNN) model is developed. The CNN model is trained with 62400 pictures of ASL of the 26 English alphabets i.e. 2400 pictures for each alphabet and on the other hand 15600 pictures are used to test the model. After training the model, the alphabets are mapped with the picture ASL so that whenever an English alphabet is given as input then the corresponding picture of ASL is viewed as the output. The schematic diagram for the first phase is shown in figure 1.

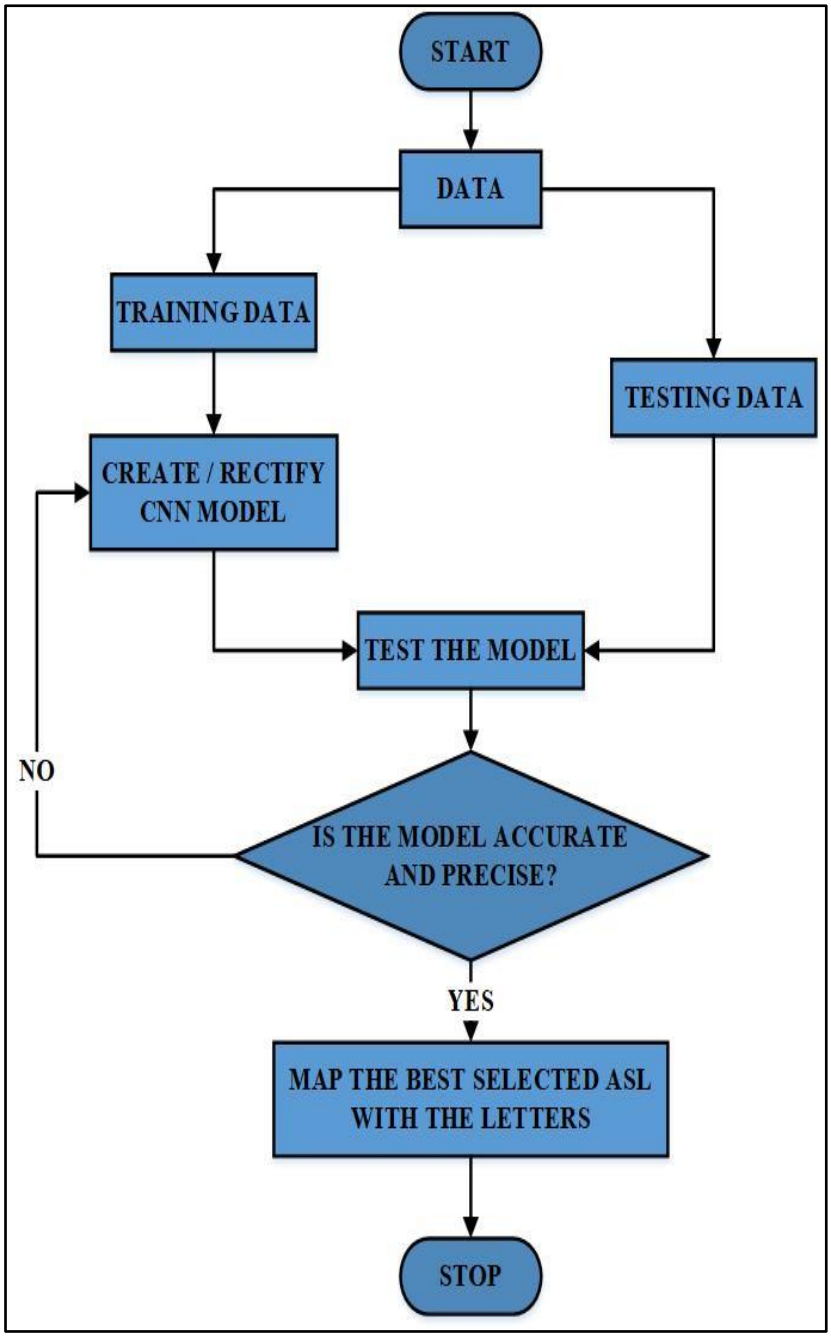


Figure 1: Schematic diagram of the first phase.

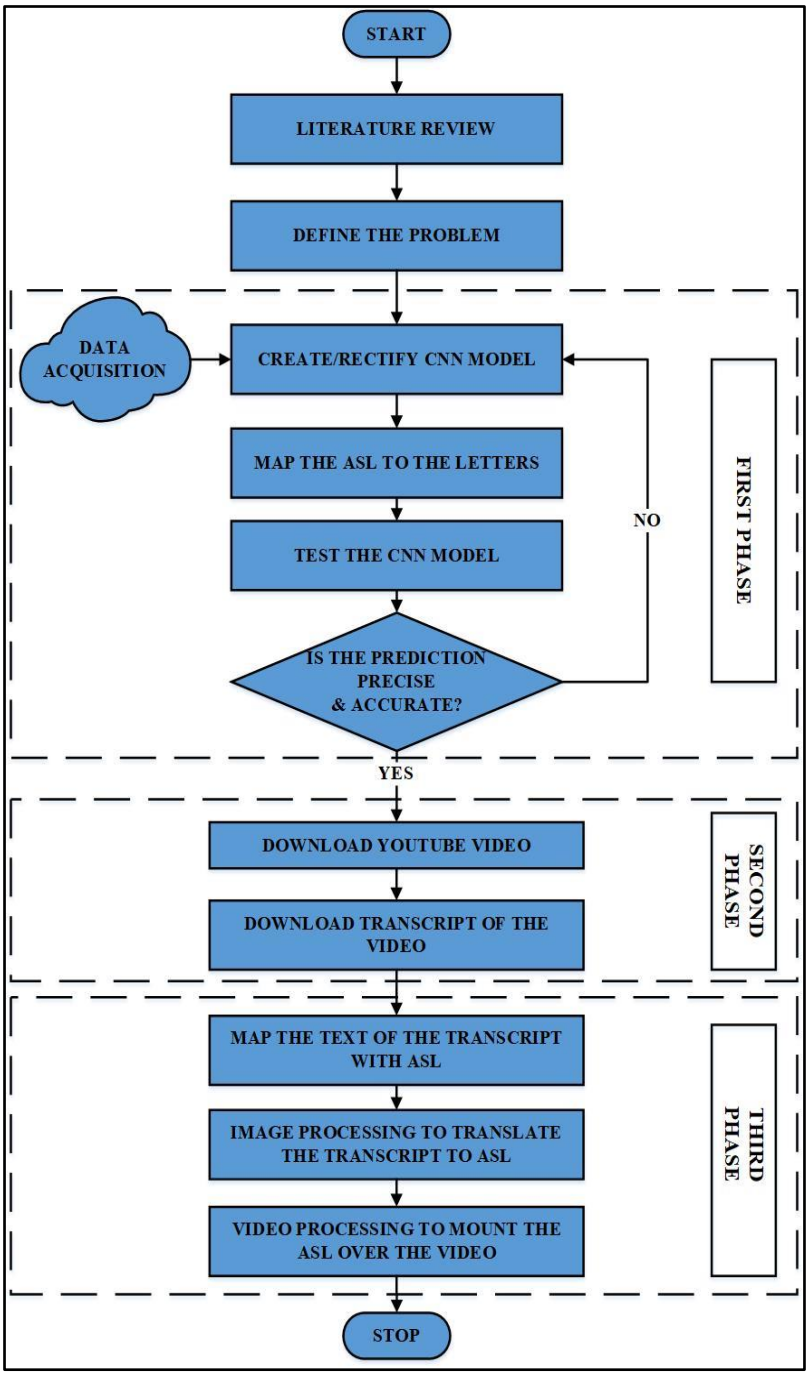


Figure 2: Flowchart of the study

The second phase of the study involves automating the download process of youtube video and its transcript by the 'youtube_transcript_api' module. The text in the downloaded transcript is then saved as text file in uppercase as the CNN model is trained only for the uppercase alphabets. The final phase of the study is converting the text of the transcript to ASL and then superimposing it on the video as per the frame rate. The third phase is executed by using the opencv and numpy modules of python. The flowchart of the study is shown in figure 2.

The paper is organized in 6 sections. Section 2 summarizes a comprehensive review of contemporary literature. Section 3 discusses the preliminary concept and the models used in the paper. Section 4

discusses the problem statement which is followed by results and discussion obtained after implementing the discussed methods on the case study in section 5. Finally, section 6 discusses the conclusion and the results obtained.

2. Review of the related works

ASL is mostly used because of its uses and at an approximation of 6,000 gestures for common words and finger spelling, recondite words or proper nouns [1]. With the increased number of researches in the direction of automation many researchers have aimed at real time ASL recognition. The first project ever to be reported for gesture recognition was in the year 1995 that used hidden Markov models [3]. However, since then, there has been much advancement made in the direction of automatic recognition of hand-gestures.

There are literatures that involve sensor-based gesture recognition of ASL including the usage of motion gloves, Kinect Sensor, image processing with cameras and leap motion controllers. In [4] an artificial neural network (ANN) model developed a 3D motion for 50 ASL words. In the paper, motion gloves were designed to recognize the ASL gestures. However, the method is time consuming and may result in imprecise calibrations caused by the wear and tear of the gloves. This drawback was reported in the literature [5 - 7]. Various factors such as sign complexities, constant finger occlusions, high interclass similarities and significant interclass variations poses a tough challenge to the task of real time ASL gesture recognition by Kinect sensors [8, 9]. Above that, the calibration of the sensory data also plays an important role in gesture recognition. In the literature, there are many studies that aim at measuring the angular positions to predict the motion gestures [10]. In [11, 12] proposed an advanced kNN methodology for the purpose of real time gesture recognition. Readers can refer to the review article on sensory gloves for sign language recognition [13 - 17]. Many papers were reviewed for conducting the study but limiting the literature review to the most recent and significant papers.

3. Methodology

Two methodologies are adopted in order to achieve the objectives in the paper. The first method involves developing a deep learning model based on CNN and the second method is video processing. At first defining the working principle and steps of CNN.

3.1. Convolution Neural Network (CNN)

CNN showed groundbreaking results over the past decade in a variety of fields related to pattern recognition; from image processing to voice recognition. The steps carried out for identifying and classifying the pictures of ASL of the letters of English alphabets by CNN are as follows:

Step 1: Acquisition of the data
Step 2: Preprocessing of the dataset
Step 3: Training of the dataset
- **Sub step a:** Performing the convolution operations
- **Sub step b:** Performing the activation operations
- **Sub step c:** Performing the pooling operations
- **Sub step d:** Layer stacking to reduce error in the model
- **Sub step e:** Classifying the outputs

Step 4: Testing the developed model

Step 5: Predicting the outputs

The steps are discussed in detail in the subsequent subsections.

3.1(a). Datasets

The dataset gathered for the study are pictures of the English alphabets in ASL. There are 26 letters in English alphabets and the letters are the different categories that the CNN model will classify. The dataset comprises 78,000 pictures of the letters of English alphabets in ASL. Figure 3 shows the pictures of the alphabet 'A' in ASL. Out of the total data, 80% of the data are randomly selected and separated as training data and the remaining 20% data are used for testing the CNN model.

Figure 3: Four pictures of alphabet 'A' in ASL from data collected

3.1(b). Preprocessing of the datasets

All the pictures have different aspect ratio, size, shape and format. In this step, the images were resized to 50*50 pixels and changed the format to Joint Photographic Experts Group (jpeg) format using OpenCV. All the images were sheared, zoomed and added 10% Gaussian noises to prevent model overfitting and enhance learning capability [18].

3.1(c). Training of the datasets

For training the network, the labeled images were fed to the model. After which the images were divided into a batch of 64. The model was trained for 100 epochs which is run for with 9 steps per epoch and 1 validation step. The steps followed for training the CNN model includes:

Convolution layer

The convolutional layer in CNN model extracts features from the image and discards the noises. Convolution operation is a mathematical process of two functions typically represented by $(f \; and \; g)$ to produce a third function (φ) that expresses how the shape of one is modified by the other. In the proposed CNN model the expression for computing the value of φ is the dot product of f and g which is given in Eq. (1).

$$\varphi = (f \, . g) \tag{1}$$

Representation of the filter matrix is shown in figure 4

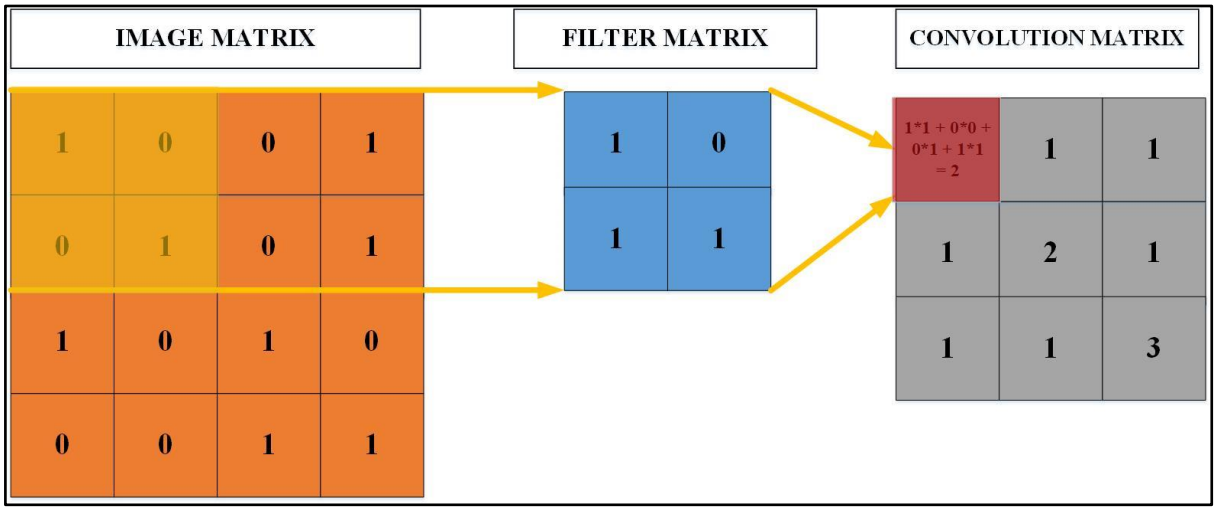


Figure 4: Pictorial representation of the convolution operation

Activation function

In neural network models, activation function transfers the input of the node to the output of the node. The activation function used in the CNN model is called a rectified linear unit (ReLU). The ReLU is mathematically expressed as:

$$y = \varphi(x) = 0, if\ x \leq 0 \quad (2a)$$

$$y = \varphi(x) = x, if\ x > 0 \quad (2b)$$

Graphical representation of ReLU is shown in figure 5.

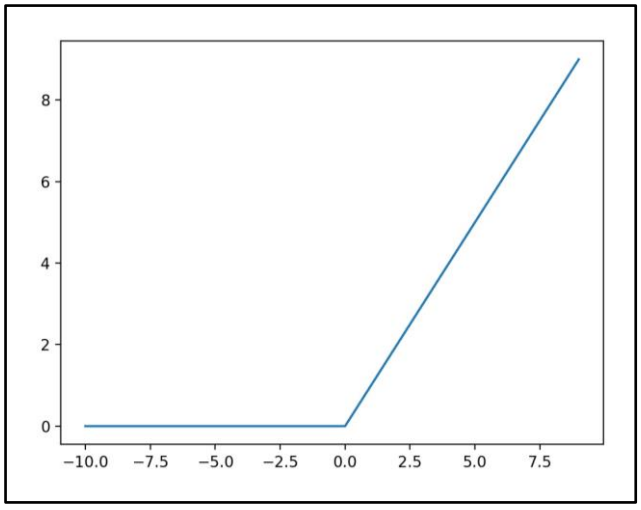


Figure 5: Graphical representation of ReLU

Pooling operations

The pooling operation reduces the dimension of the feature maps which in turn reduces the number of parameters to learn and the amount of computation performed in the network. The importance of the pooling layer is to summarize the feature present in a region of the feature map generated by the

convolution layer. So, further operations are performed on summarized features instead of precisely positioned features generated by the convolution layer. This makes the model more robust to variations in the position of the features in the input image.

Layer stacking operations

In layer stacking operation, the convolution operation, activation and pooling operation is repeated until the output obtained is a minimized matrix of the input image.

Fully connected layer

The Fully Connected (FC) Layer consists of neurons that are fully connected with the neurons from the previous layers. The FC layer predicts the output or the label of the input class. In case of multi-class problems, different activation functions are used to classify the label of the inputs. In this study, SOFTMAX activation function is used to classify the label. Hence, it has an output dimension of [1×1×M] where M is the number of classes or labels used for classification

Classification

Classification is a process related to categorization. In the study, the CNN model classifies the letters of English alphabets.

3.1(d). Testing of the datasets

To test the performance of the CNN model 20% of the dataset i.e. 15,600 pictures of the ASL of the corresponding letters of English alphabets. The 20% of the data are segregated from the training data using the train_test_split() function. The output label of the test images obtained from the CNN model is compared to the actual label of the images.

3.1(e). Prediction

Prediction of CNN model is its ability to detect and classify the different pictures of ASL of the letters of English alphabets. Figure 6 shows the diagrammatic representation of the proposed CNN model.

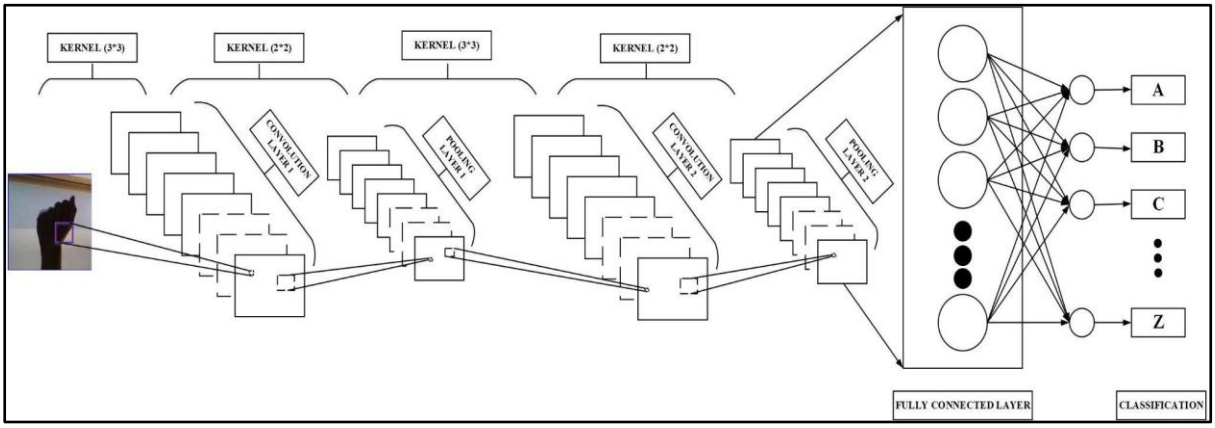


Figure 6: Diagrammatic representation of the proposed CNN model.

3.2. Image and Video processing

Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. On the other hand, video processing in electronics is a signal processing which often employs video filters and where the input and output signals are video files or video streams. In this study, image processing is used to translate the text of the transcript into the letters of English alphabets by concatenating the pictures of ASL. THE video processing is used to mount the resultant picture over the downloaded video.

3.3. Steps for the integrated methodology

The steps for the integrated methodology are as follows:

Step 1: Collect data.
Step 2: Create the CNN model.
Step 3: Map the most appropriate picture of the ASL with the letter
Step 4: Download youtube video.
Step 5: Download the transcript and timestamps of the video.
Step 6: Save the text of the transcript in upper case as all the training of the alphabets are done in upper case.
Step 7: Call the mapped ASL image for the corresponding letter.
Step 8: Concatenate the images called in step 7. This is the ASL based subtitles
Step 9: Run the video.
Step 10: Match the time stamps of the video and synchronize the ASL based subtitles

4. Case study

The primary focus of the present study is to create ASL based subtitles for the youtube videos. According to an estimate by the World Health Organization about 5% of the world's population are deaf [19] and they use SL for communication. It is often observed that the youtube videos, although the subtitles are available in English or other native languages, do not have any SL based subtitles. Therefore an attempt is made in this paper to develop ASL based subtitles for the youtube videos.

5. Results and discussions

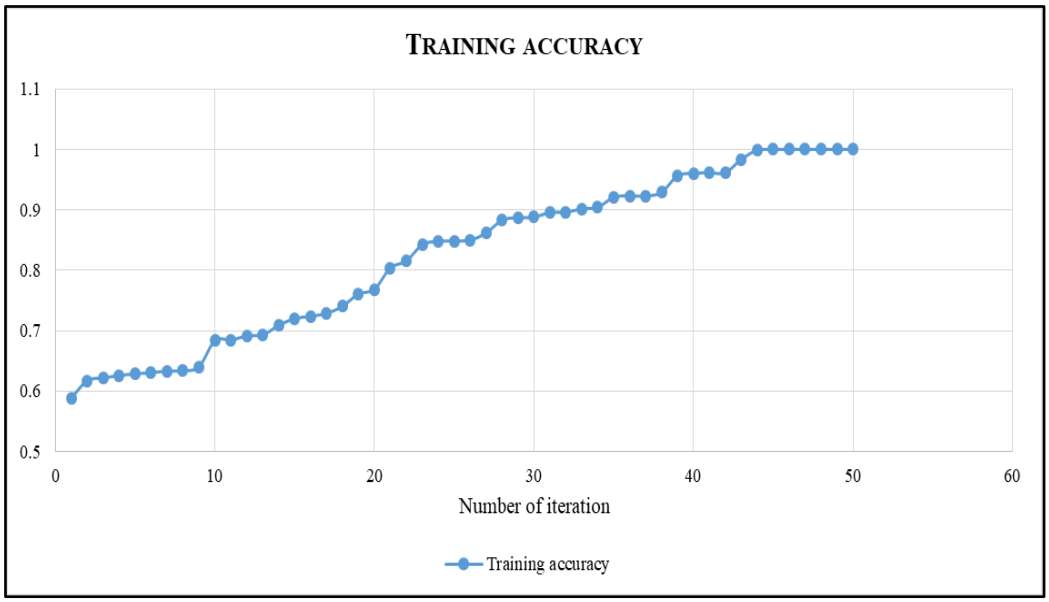
In this step the result obtained from application of the proposed method in the problem is discussed. In the study the methodology is coded in Python 3.8.5 and ran on a 64-bit windows 10 system with 8GB RAM and i5, 1.6GHz processor. The different python libraries required to carry out the analysis are listed in table 1.

Table 1: Different libraries imported

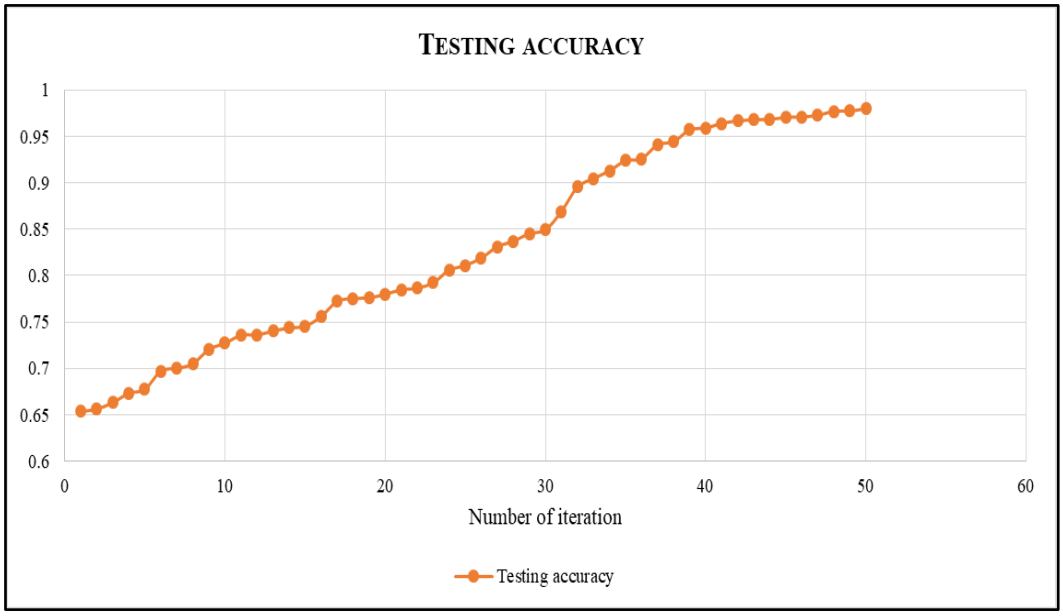
Sl. no.	Library imported	Usefulness	Sl. no.	Library imported	Usefulness
1	Operating system (os)	For using os dependent functionality.	**9**	Selenium	Web-based automation tool
2	Numpy	Used for array operations	**10**	webdriver_manager	Automates the browser setup in the Selenium code
3	Torch	Used for scientific computing framework	**11**	Opencv	Image and video processing
4	Glob	Used for finding pathnames	**12**	pytube	Downloading youtube videos
5	Torchvision	Used for machine learning framework	**13**	shutil	Used to copy the content of the source file to the destination file
6	Pathlib	Provides an object API for working with files and directories	**14**	math	Used for mathematical functions
7	youtube_transcript_api	Downloads transcripts/subtitles for YouTube videos.	**15**	datetime	Classes for manipulating date and time
8	matplotlib	Used for creating static, animated, and interactive visualizations	—	—-	—

The CNN model is executed for 50 iterations and the performance of the CNN model developed in the study is computed by plotting the accuracy of the training and testing data. The accuracy curve for the training and testing data is shown in figure 7.

From figure 7, it is observed that the CNN model developed the training accuracy of the proposed CNN model increased from 58.86% in the first epoch to 99.982% in the 50th epoch. On the other hand the training accuracy of the CNN model increased from 65.43% in the first epoch to 98% in the 50th epoch. The CNN model is capable of accurate and precise prediction.



(a)



(b)

Figure 7: (a) Training and (b) Testing accuracy of the developed CNN model

After following the steps of the integrated methodology as discussed in section 3, the model is tested on a random video which was downloaded from youtube. The output of the frame at 2:26 minute is extracted and saved in jpg format first with subtitles and next without subtitles. The subtitles at the 2.26 min and the ASL for the phrase is shown in figure 8.

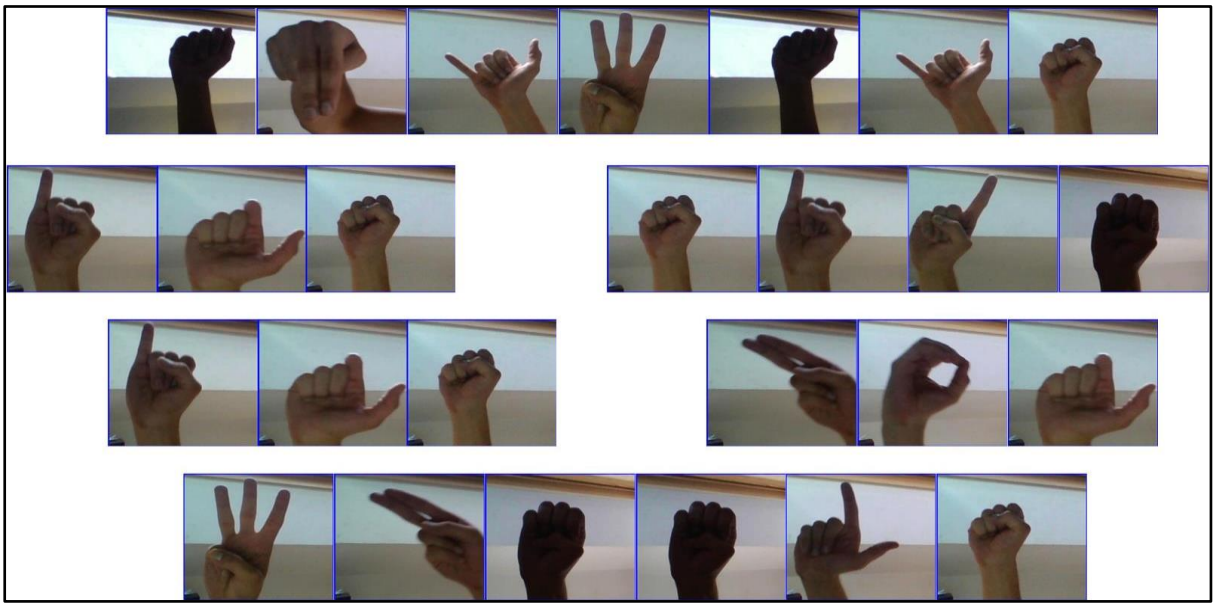


Figure 8: Translation of the subtitles in 2.26 min of the video in ASL

The frame of the 2.26 minute of the youtube video with normal and ASL based subtitles is shown in figure 9.

(a)

(b)

Figure 9: The frame of the 2.26 minute of the youtube video with normal and ASL based subtitles

6. Conclusion

The comprehensive intention of the present study is to develop a model that can automate the conversion of text in the transcript of youtube videos to ASL. Therefore, the study proposes a framework that can help to achieve the aim of the present study. The content for developing the framework is divided into three phases. The first phase involves the development of a torch based CNN model to map the letters of English alphabets into ASL. The CNN model showed training and testing accuracy of 99.982% and 98% respectively. The second phase involves the process of downloading the youtube video and its transcript followed by conversion of text into uppercase. The final phase of the paper involves calling the most appropriate ASL image of the English alphabets and concatenating them according to the letters in the word and then mounting it over the video. The strength of a model lies in its ability to be applied in a practical problem. Therefore, the proposed integrated method is applied to extract a random video from youtube. The result obtained from applying the integrated method to the video showed that the result obtained is highly satisfactory and can be applied for future use.

Acknowledgement

I would like to express my heartfelt gratitude to OnMyOwnTechnology for extending their help in carrying out the particular project. I shall remain ever grateful for their help and generosity.

Conflict of interests

The authors would like to declare that there is no funding received in any form from private, governmental and semi-governmental agencies in carrying out this project.

References

1. Starner, T., Weaver, J., & Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video. IEEE Transactions on pattern analysis and machine intelligence, 20(12), 1371-1375.
2. Mitchell, R. E., Young, T. A., Bachelda, B., & Karchmer, M. A. (2006). How many people use ASL in the United States? Why estimates need updating. Sign Language Studies, 6(3), 306-335.
3. Starner, T. E. (1995). Visual Recognition of American Sign Language Using Hidden Markov Models. Massachusetts Inst Of Tech Cambridge Dept Of Brain And Cognitive Sciences.
4. Oz, C., & Leu, M. C. (2011). American sign language word recognition with a sensory glove using artificial neural networks. Engineering Applications of Artificial Intelligence, 24(7), 1204-1213.
5. Oz, C., & Leu, M. C. (2007). Linguistic properties based on American Sign Language isolated word recognition with artificial neural networks using a sensory glove and motion tracker. Neurocomputing, 70(16-18), 2891–2901. https://doi.org/10.1016/j.neucom.2006.04.016
6. Huenerfauth, M., & Lu, P. (2010). Accurate and accessible motion-capture glove calibration for sign language data collection. ACM Transactions on Accessible Computing (TACCESS), 3(1), 1-32.
7. Luzanin, O., & Plancak, M. (2014). Hand gesture recognition using low-budget data glove and cluster-trained probabilistic neural network. *Assembly Automation*.

8. Sun, C., Zhang, T., Bao, B. K., Xu, C., & Mei, T. (2013). Discriminative exemplar coding for sign language recognition with kinect. IEEE Transactions on Cybernetics, 43(5), 1418-1428.
9. Tao, W., Leu, M. C., & Yin, Z. (2018). American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion. Engineering Applications of Artificial Intelligence, 76, 202-213.
10. Fujiwara, E., dos Santos, M. F. M., & Suzuki, C. K. (2014). Flexible optical fibre bending transducer for application in glove-based sensors. IEEE Sensors Journal, 14(10), 3631-3636.
11. Tubaiz, N., Shanableh, T., & Assaleh, K. (2015). Glove-based continuous Arabic sign language recognition in user-dependent mode. IEEE Transactions on Human-Machine Systems, 45(4), 526-533.
12. Aly, W., Aly, S., & Almotairi, S. (2019). User-independent American sign language alphabet recognition based on depth image and PCANet features. IEEE Access, 7, 123138-123150.
13. Lee, B. G., & Lee, S. M. (2017). Smart wearable hand device for sign language interpretation system with sensors fusion. IEEE Sensors Journal, 18(3), 1224-1232.
14. Paudyal, P., Lee, J., Banerjee, A., & Gupta, S. K. (2019). A comparison of techniques for sign language alphabet recognition using armband wearables. ACM Transactions on Interactive Intelligent Systems (TiiS), 9(2-3), 1-26.
15. Wu, J., Sun, L., & Jafari, R. (2016). A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors. IEEE journal of biomedical and health informatics, 20(5), 1281-1290.
16. Wu, J., & Jafari, R. (2017). Wearable Computers for Sign Language Recognition. In Handbook of Large-Scale Distributed Computing in Smart Healthcare (pp. 379-401). Springer, Cham.
17. Wu, J., Tian, Z., Sun, L., Estevez, L., & Jafari, R. (2015, June). Real-time American sign language recognition using wrist-worn motion and surface EMG sensors. In 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN) (pp. 1-6). IEEE.
18. Gu, S., Pednekar, M., & Slater, R. (2019). Improve image classification using data augmentation and neural networks. *SMU Data Science Review*, *2*(2), 1.
19. Deafness and hearing loss, retrieved from https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss on 30th April, 2022.