Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, March : 2023

A NEW APPROACH ON CNN BASED SIGN LANGUAGE TO SPEECH TRANSLATION

1D.S.Chandra Mouli, 2K.Durga Mahesh, 3B.Hema, 4B.Divya, 5M.Jyothirmai, 6T.Gowri Sankar

1 Assistant Professor in Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences(Autonomous), Rajampet, Andhra Pradesh, India.

2,3,4,5,6Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences(Autonomous), Rajampet, Andhra Pradesh, India.

Abstract: Despite the fact that a tiny proportion of the world's population suffers from speech and hearing difficulties, sign language is not a commonly spoken language. In today's oral culture, gesture-based languages are not particularly widespread among the general population. Despite the fact that sign language has developed significantly over the years and is used in many different regions of the globe language. Hence, improvements in technology may enhance communication between the speaking and non-speaking worlds. It is crucial to have a real-time application that permits sign language communications between individuals who can sign and those who cannot. The issue addressed in the problem statement of this article has an application as its solution. To create a real-time reacting application that would facilitate a live conversation between a speaking and non-speaking individual, live video inputs that would subsequently be converted into speech are required. In this study, text-to-speech translation and convolutional neural networks (CNN) are proposed as complementing technologies. The CNN algorithm is used to convert the input text into speech.

Keywords: Text-to-speech algorithms, neural networks, and real-time gesture recognition for Indian Sign Language.

I. INTRODUCTION

While being widely used by speech and hearing-impaired people, sign languages are not very popular in the speaking world. Communication between the two communities is hampered as a result. These languages are challenging for the typical individual to grasp since they are recognised natural languages with distinctive lexicons and grammars.

American Sign Language (ASL) and British Sign Language are the two most popular sign languages, and they each have regional variations (BSL). Indian Sign Language is logically the name given to the sign language utilised in India (ISL). Despite its extensive usage throughout the country, the spoken populace clearly has little to no knowledge of this signing vernacular. The Indian government has even supported the establishment of an online dictionary, despite the fact that there already a tonne of learning materials for this particular style of sign language accessible.

The second version of the ISL Dictionary, which included 1000 new words, was published in February 2019 [1]. The dictionary's first edition, which has 3000 words, was published in March 2018.

This programme is offered by the ISLRTC's Department of Empowerment of People with Disabilities (DEPwD) of the Ministry of Social Justice & Empowerment [2]. It is acknowledged that these minority continue to face prejudice in today's society [3]. To increase awareness, the general public must have access to the learning resources for sign language.

Around 50 lakh people in India alone suffer hearing impairments, according to the 2011 Census [4], which includes 19% [5] of the population. Speech issues affect over 20 lakh people, or 7% of the population [4]. Artificial intelligence-assisted technology is being used to bridge this communication gap and raise the standard of living for these underrepresented groups.

Many technologies have made it simpler to translate sign languages into spoken languages and the other way around. Despite efforts, technologies for daily usage have not yet been created and made generally accessible.

The available sign to text or sign to voice translators are not sufficiently sophisticated to meet the needs of the target audiences or integrate into daily life. In contrast, the design of sign to voice or text translators is rather straightforward.

Thus, there is a pressing need for study into the development of technology that may be adapted to the everyday needs of people with speech and hearing impairments. It is necessary to find a solution that enables non-speaking and non-hearing cultures to interact with hearing and speaking individuals in natural ways during ordinary interactions. Or to put it another way, we need to develop a technology or interface that will make it possible for anybody to comprehend spoken sign language when it is used to interact with them.

II. LITERATURE SURVEY

The translation of sign languages has been made possible by a variety of technological advancements. These technologies range from software-only applications to some that utilize hardware. Systems with glove-like [6] extensions that the signer can wear to enable gesture capture, identification, and ultimately translation have been suggested and built. With increased accuracy, these gloves assist in capturing the signer's hand movements, which are then translated using the right algorithms. Such devices are limited in that they only record hand motions, not facial expressions or body language, which is a flaw. Pictures or videos are therefore taken as inputs as an alternative to tackle this problem. Using either mobile devices or specialist devices with built-in cameras, these applications capture user input through cameras. As an illustration, a team from Tamil Nadu developed a smartphone application that makes use of video inputs [7]. Taking video inputs while wearing a camera on a cap is one way to collect inputs for this application. This cap, which has been modified, would be worn by the signer, and it would record

all of her or his actions from the top view. Hand gestures, facial emotions, and body language all play a part in sign language, as was previously said. These kinds of models narrow their attention to handbased movements and exclude other aspects of signing's aesthetics. Within the confines of their research, these models do, however, deliver results with varied degrees of accuracy. In order to account for body language, the method should be modified to accept video inputs that record the signers' facial expressions in addition to their hand motions. In addition to making, it possible to work on the data and recognise the signs more efficiently, this would also allow for the capture of the other signing-related factors. The range of parameters involved in signing is one of the main obstacles addressed in sign language translation. Other different factors need to be taken into account in addition to the fundamental aesthetics of signing in any of the sign languages. The signature of the same message has unique differences. As a result of the variances in their unique actions and body languages, two people signing the same word or sentence might produce different signs. Additionally, each person's hands vary physically and structurally. When providing signals as input to an algorithm or programmer, these criteria matter. The training database determines how well an algorithm performs. In order to capture the widest range of variations, it is important to ensure that the training database contains a diversity of signs from different signers. Following that, image processing and computer vision techniques like skin detection and Canny Edge detection are used to process the supplied photos or videos. These methods can be used explicitly or in conjunction with the current gesture recognition algorithms. A variety of algorithms have traditionally been used for the generation of such translators depending on the nature of the model that has been proposed. SURF (Speeded-Up Robust Features), SIFT (Scale Invariant Feature Transformation) [8], and neural networks have all been used, depending on the needs in each situation. Artificial neural networks (ANN), often known as neural networks, are a type of computing network that is based on biomimicry, or the imitation of biological processes like the connections that naturally exist between neurons in animal brains. These are connectionist systems as a result. Deep Neural Networks are a subcategory of Artificial Neural Networks (ANN) that can be modified to support deep learning. The implementation of deep neural networks can be done in several different ways. One example of this is the use of convolutional neural networks (CNN). A completely linked system that uses convolutions in one or more layers of neuron connectivity is referred to as a CNN, also known as a Space Invariant Artificial Neural Network (SIANN). ConvNets, also known as CNNs, are frequently used in the areas of facial recognition and object detection, among other things, because of their great value in picture recognition and classification. [9] The results of the application of the aforementioned method ought to provide a text output of the input signs. In order for this output to be useful for this system, it would then need to be transformed to audio. Thus, the final component of this equation would be text to speech conversion. The conversion of text to speech is facilitated by a number of APIs and libraries. These packages, which are typically open source and available for many different languages, provide for implementation flexibility by doing away with the need for a particular language. The Google Test-to-Speech (gTTS) [10] is one of many Test-to-Speech translating libraries available for Python. This library is online, therefore any programme using it would need to have an active internet connection for it to work properly in real time. The library can't operate in offline mode, but it is very programmable

and adaptable and may be used to store the outputs produced in independent audio files. To create technology that will allow the translation of Indian Sign Language is the main reason behind this study (ISL). ISL will be the focus of the project because there hasn't been much work done on it, even though the model is scalable and can be used for multiple Sign Language dialects with the right training dataset. For the deaf and mute community in India, the ideal situation would be to create a user-friendly, widely accessible application. Although there has undoubtedly been progress in this regard, real-time translation offers the potential for even greater translation accuracy.

III. PROPOSED SYSTEM

In order to translate the signs to spoken English, this translation explicitly considers the Indian Sign Language (ISL). The following steps are part of the strategy used to put this paradigm into practice:

1. The creation or compilation of a database of signs in a selected sign language.

2. By using and implementing a neural network based algorithm, sign recognition from input feed is achieved.

3. Training the model and improving translation accuracy through processing utilising classification and machine learning techniques.

4. Creation of an input sign text translation

5. Transforming Text into Speech A testing and training phase to hone the model would come after the aforementioned implementation.

A. Flow Diagram

The flowchart details the procedures that must be taken to complete

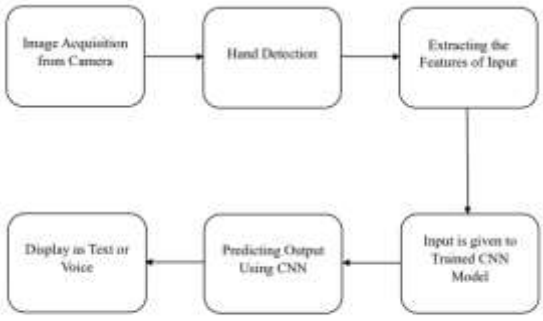

Figure 1 Project Flow Chart outlining the project's goals.

Below is a more thorough explanation of these steps:

1.Image Acquisition

With the help of the web camera, the gestures are recorded. To record the entire signing session, an OpenCV video stream is used. The frames are pulled from the stream and processed as grayscale images with a 50x50 pixel size. Due of the uniform size of the dataset across the board, this dimension stays constant throughout the project.

2. Finding the hands,**Extracting the Features**

Searching for hand motions in the collected photos. Prior to feeding the image to the model to get the prediction, this is a step-in preprocessing. A stronger emphasis is placed on the gesture-filled sections. By many folds, this raises the likelihood of prediction.

3. Hand Posture Recognition

The keras CNN model receives the preprocessed pictures. The projected label is produced by the trained model. Every label for a gesture has a probability associated with it. The anticipated label is assumed to be the one with the highest likelihood.

4. Show as Text and Speech

The model converts identified gestures from gestures to words. The pyttsx3 library is used to translate the identified words into the matching speech. It is a straightforward workaround, but the text to speech outcome is a priceless feature because it simulates a vocal dialogue.

B. Convolutional Neural Network for Detection

Neural networks in the CNN class are quite effective in addressing computer vision issues. Inspiration for their work came from how our brains actually perceive vision, namely in the visual cortex. To enable the detection of a given feature, they employ a Image Acquisition from Camera Extracting the Features of Input Hand Detection Predicting Output Using CNN Display as Text or Voice Input is given to Trained CNN Model filter or kernel to iteratively scan over all of the image's pixel values and do calculations by establishing the proper weights.

Convolution, max pooling, flatten, dense, dropout, and a fully connected neural network layer are just a few of the layers that the CNN is made up of. An extremely effective tool that can recognise features in an image is created by these layers working together. Beginning with the detection of simple features, higher-level features of increasing complexity are eventually picked up by the initial layers.

C. The CNN Architecture functioning

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, March : 2023

In this project, there are 11 layers in the CNN model. Convolutional layers number three. A grayscale image with a 50x50 pixel size can be submitted to the first convolutional layer, which is in charge of identifying low level features like lines. This layer contains 16 filters of size 2*2, which produces an activation map of size 49*49 for each of the filters and an output of size 49*49*16. Additionally, a rectifier linear unit (relu) layer is added to the map, replacing all negative values with 0. By only taking into account maximum values in 2*2 sections of the map, the activation is reduced to 25*25 by applying a maxpooling layer.

The likelihood of finding the target feature rises as a result of this phase. A second convolutional layer is added after that. It is in charge of locating characteristics like angles and curves. The output of this layer, which includes 32 filters of size 3*3, is equivalent to 23*23*32 due to the 23*23*23 activation map that is produced. The activation map is further reduced to 8*8*32 by a maxpooling layer, which looks for the highest values in 3*3 sections of the map. High level elements like movements and forms are recognised using a third convolutional layer. The input is reduced to an output of size 4*4*64 by 64 filters of size 5*5.

The map becomes 1*1*64 after a maxpooling layer. The map is compressed into a 1D array with a length of 64. The map grows to a 128-element array thanks to a dense layer. To lessen overfitting, a dropout layer removes arbitrary map elements. The map is ultimately reduced to an array of 44 items, which stand in for the number of classes, by the dense layer.

The chance of prediction assigned to each class is corresponding. As the projected gesture, the class with the highest probability is shown.

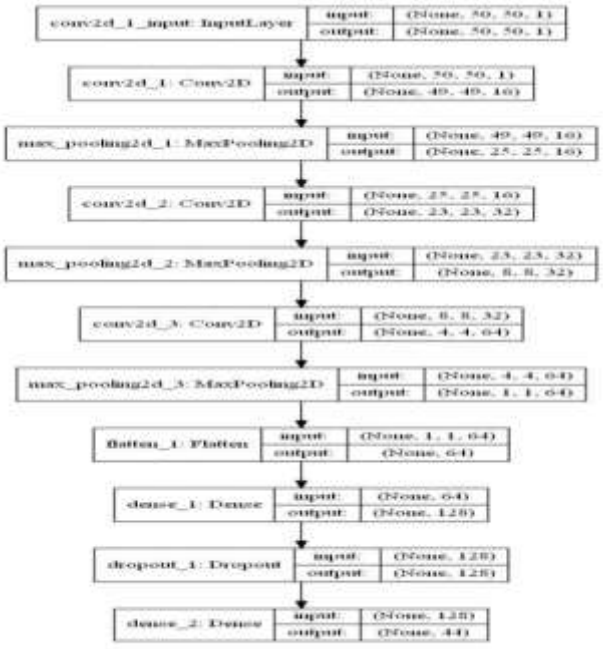

Figure 2 The project's CNN architecture

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, March : 2023

D. Recognition of Alphabets and number

We employed the Gaussian historical past subtraction technique, which involved updating each historical pixel using a combination of K Gaussian set distributions, to find the bounding boxes of diverse objects (k from 3 to 5). The colours that last longer and are more static are potentially historical ones. We create a square bounding field for those varying pixels. A Convolutional NN model has been developed utilising those photographs after collecting all the signs and historical data to separate the gesture symptoms and indicators from the past. These function maps describe how the CNN is able to comprehend the typical unexposed structures of some of the sign indicators in the training set and is subsequently able to differentiate all the signs.

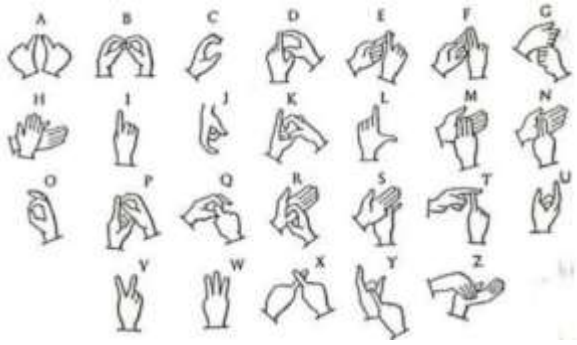

Figure 3:The ISL Alphabet Symbols that will be included in the training data

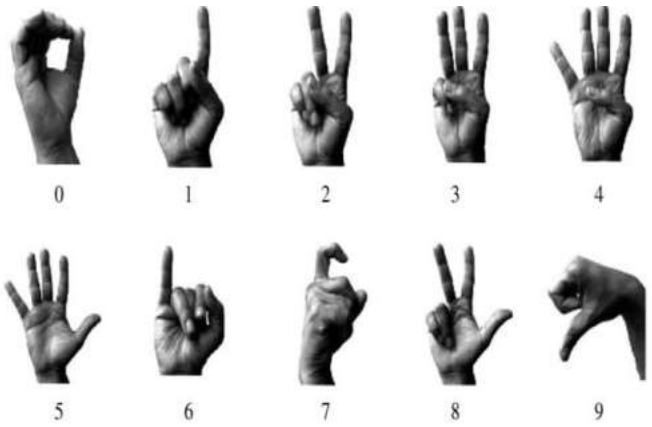

Figure 4 :The Symbols for ISL Numbers that will be in the training data

E. Algorithm

Real-time sign language to text conversion algorithm, begin

S1: In accordance with the skin tone and illumination, change the hand histogram.

S2: To increase the dataset and hence lessen overfitting, use data augmentation.

S3: Create train, test, and validation data sets from the dataset.

S4: To fit the dataset, train the CNN model.

S5: The accuracy, error, and confusion matrices should be included in the model report that is generated.

S6: Execute the prediction file, which combines gesture predictions into words and displays them as text while relaying vocal output.

IV. RESULTS

Figure 5:Screenshot of the result obtained for alphabet A

Figure 6: Screenshot of the result obtained for number 8

V. CONCLUSION

The project is a simple example of how CNN may be used to resolve computer vision issues with a very high degree of accuracy. A 98.5% accurate finger spelling sign language translator is available. The project may be extended to incorporate other sign languages by developing the relevant dataset and training the CNN. Since sign languages are spoken more in context than as finger typing languages, a component of the Sign Language translation issue may be handled by the project. The current project only supports ISL, however it may be developed to support more native sign languages given enough dataset and training. Future modifications to this project may be made in a variety of ways. To make it easy for consumers to access the project, it may be created as a web or mobile application, for instance. Despite the employment of a finger spelling translator in this study, sign languages are also spoken in context, with each gesture perhaps designating an object or verb. This form of contextual signing would thus need more processing power and natural language processing to recognise (NLP).

REFERENCES

[1] Indian Sign Language Dictionary Launch, 2019 [Online] Available: http://www.islrtc.nic.in/isldictionary-launch.

[2] 2nd Edition of Indian Sign Language Launched, 2019 [Online] Available: https://currentaffairs.gktoday.in/2nd-edition- indian-sign-language-dictionary-launched02201966449.html

[3] NPR, A Mom Fights to Get An Education For Her Deaf Daughters, 2018 [Online] Available:https://www.npr.org/sections/goatsandso da/2018/01/14/575921716/a-mom-fights-to-getan-education-for-her-deaf-daughters

[4] Department of Empowered Persons With Disabilities (Divyangjan), State UT Wise Persons with Disability: Number of Disabled Persons Disability Wise as per Census 2011, 2015 [Online] Available: http://disabilityaffairs.gov.in/content/page/state-utwise-persons.php

[5] Disabled Persons in India: A Statistical Profile 2016, 2016 [Online] Available: http://mospi.nic.in/sites/default/files/publication_re ports/Disabled_persons_in_India_2016.pdf

[6] D. Sharma, D. Verma, P. Khetarpal, 'Labview Based Sign Language Trainer Cum Portable Display Unit For The Speech Impaired', in 2015 Annual IEEE India Conference (INDICON), 2015

[7] Y. Madhuri, G. Anitha, M. Anburanjan, 'VisionBased Sign Language Translation Device', in 2013 International Conference on Information Communication and Embedded Systems (ICICES), 2013

[8] SIFT:Theory and Practice [Online] Available: http://aishack.in/tutorials/sift-scaleinvariant-feature-transform-introduction/

[9] Understanding of Convolutional Neural Network (CNN) — Deep Learning, 2018 [Online] Available: https://medium.com/@RaghavPrabhu/understandi ng-of-convolutional-neural-network-cnn-deeplearning-99760835f148

[10] gTTS.PyPI, gTTS 2.0.4 [Online] Available: https://pypi.org/project/gTTS/

[11] Indian Sign Language Recognition System [Online] Available: https://github.com/abhishekdudhal/Indian-SignLanguage-Recognition-System

[12] ImageNet: VGGNet, ResNet, Inception, and Xception with Keras, 2017 [Online] Available: https://www.pyimagesearch.com/2017/03/20/imag enet-vggnet-resnet-inception-xception-keras/
[13]Convert Text to Speech in Python [Online] Available: https://www.geeksforgeeks.org/converttext-speech-python/

[14] Text to Speech [Online] Available: https://pythonprogramminglanguage.com/text-tospeech/

[15] J. Singha, K. Das, 'Recognition of Indian Sign Language in Live Video', in International Journal of Computer Applications (0975 – 8887) Volume 70– No.19, May 2013

[16] P. V. V. Kishore, P. Rajesh Kumar, 'A Video Based Indian Sign Language Recognition System (INSLR) Using Wavelet Transform and Fuzzy Logic', in IACSIT International Journal of Engineering and Technology, Vol. 4, No. 5, October 2012