Sign Language Recognition and Text-to-Speech Translation

MSc Research Project
Data Analytics

Sumitra Sambhaji Thorat
Student ID: x20248521

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sumitra Sambhaji Thorat
Student ID:	x20248521
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Athanasios Staikopoulos
Final Submission Date:	01/02/2023
Project Title:	Sign Language Recognition and Text-to-Speech Translation
Word Count:	4,820
Page Count:	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sign Language Recognition and Text-to-Speech Translation

Sumitra Sambhaji Thorat
x20248521

Abstract

Communication is the most important aspect while conveying messages. The ability to listen and speak plays an important role while communicating. There are people who have difficulty in listening or speaking. As humans use verbal language, conversing with specially-abled people become a challenge. Because of the mentioned reason these people use sign language to communicate and everyone is not familiar with sign language. The unfamiliarity with sign language becomes a barrier while communicating. This research will assist the people who are unable to speak and listen to converse with normal people by using sign language recognition and converting these in text or speech. This is implemented using American and Indian sign language to reduce language barrier. Convolution Neural Network (CNN) was used for Indian Sign Language (ISL) and American Sign Language (ASL). In both dataset Image Augmentation was performed and text to speech was converted using Google Text to Speech library from python. The ISL achieved the accuracy of 99% and the loss observed was 0.5. The accuracy of ASL was observed to be 96% and the loss was 0.11.

1 Introduction

1.1 Background

While communicating, familiarity with the language is important. Whenever someone wants to express or say something, they communicate in a language with which they are familiar or comfortable. Language is crucial when it comes to communicating. If someone is not familiar with the language in which the other person is trying to communicate, it is like gibberish for the person who is listening to it. There are people in world who are not able to listen or speak or have difficulty in it. In 2015, the author (Rastogi et al.; 2015) has mentioned that over 300 million people who have issues with hearing and 1 million people who have issues with speaking. These figures are from the World Health Organisation (WHO). Further investigating into the numbers, WHO has mentioned that globally there are 1.5 billion people with hearing disability and this could grow over 2.5 billion by 2050[1]. Language becomes a barrier during communication with these people. They make use of sign language to communicate. The person who is not familiar with sign language finds it difficult to communicate with the people who are unable to speak and listen.

[1] https://www.who.int/health-topics/hearing-loss#tab=tab_1

1.2 Motivation

Sign language is language which is formed with the movement of the hands. There are different sign languages like American sign language, British sign language, Irish sign language, Indian sign language etc., the person who can communicate in American sign language may not understand British sign language. When communication becomes a barrier, the people with hearing and speaking issues may feel excluded. This can also hamper the growth of the country. According to recent studies, these people have high unemployment and low education rates. This leads to additional health and educational support and less productivity of the respective country[2] . This research study will assist individuals with hearing and speaking disabilities to communicate effectively. The language will not become a barrier while expressing themselves. As the world is accepting inclusion and diversity, the people who are unable to speak and listen feels neglected or left out as there are not much people who are familiar with the sign language. Also, over the world there are 300 different sign languages[3], it is not possible that everyone can know every language. This research is based on Indian and American sign language.

Below Figure 1 shows how the signs change in Indian and American Sign Language. It can also be seen that Indian sign language uses both hands, whereas American sign language uses one hand.

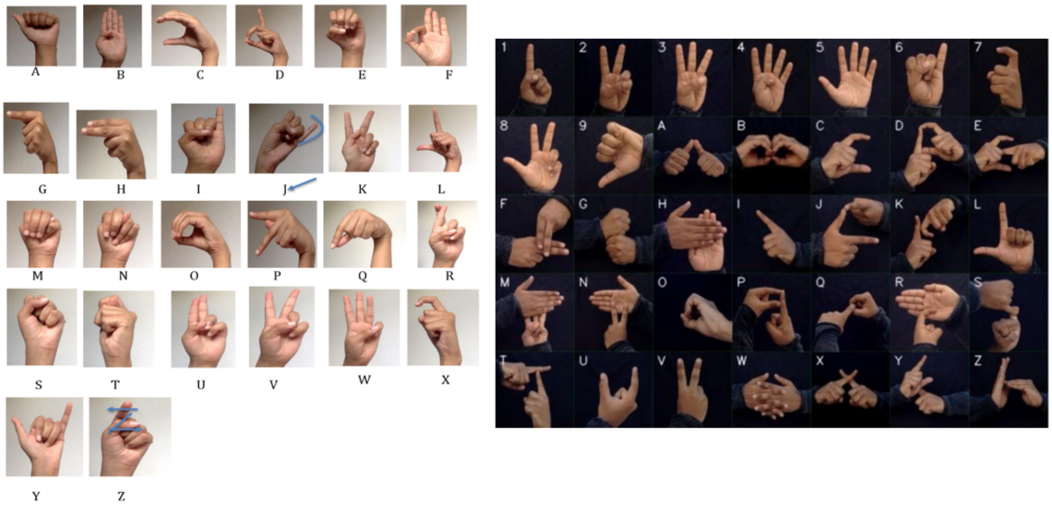


Figure 1: ASL & ISL

1.3 Limitation

Various studies have been conducted in similar fields for the recognition of hand movements and their gestures. The complexity of tracking the moving hand compared to the static hand led them to use the gloves for tracking the moving hands. These gloves consist of sensors and many electronic components to help them recognise the words. These gloves are innovative solution but they are expensive, complex[4], sometimes restrictive and sometimes unclear (Kumar et al.; 2018).The vision-based sign language was implemented

[2] https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[3] https://en.wikipedia.org/wiki/List_of_sign_languages

[4] https://newatlas.com/sign-language-translate-glove/50474/

using Fuzzy Cluster Mean (FCM) was time-consuming. Many researchers who worked on vision based recognition had small datasets and these studies have been conducted on single sign language.

1.4 Research Question and Objectives

To overcome the above stated limitations, this research has proposed a solution where American and Indian Sign language has been integrated and a system has been built where people from these regions knowing sign language can communicate. This study provides a solution to the below research question:

How efficiently can Convolution Neural Network identify and translate American and Indian sign language?

The main objective of this study is to have fluent conversations between people who are unable to speak and listen. This can make people who have difficulty speaking and listening feel included in society and comfortable while conversing. This project will translate the sign languages of America and India into text and speech. This will result in having conversation without difficulty between normal person and the person with the special abilities as well as the person from various these two regions of the world who are not able to speak and listen. The project has implemented the following:

- Extraction, cleaning, analysing and pre-processing of the image dataset
- Building model to identify and translate the signs in text and speech
- Evaluate the model.

1.5 Report Synopsis

This research paper has been divided into sections. Following are the sections discussed throughout the research paper: Section 2 covers the research work conducted in previous years, Section 3 will discuss the methodology used to conduct this study this section will discuss the pre-processing cleaning and modelling of data, Section 4 will cover the design specification, Section 5 will discuss the Implementation of the research, Section 6 will evaluate the models, Section 6.3 is discussions and Section 7 is Conclusion and Future Works.

2 Related Work

Around the world, there are various languages in which people communicate. Likewise, there are several sign languages which belong to specific regions. These sign languages help the people with hearing and speaking issue to communicate effectively. For example, India uses Indian Sign Language, America has American Sign Language, then there is British, German, Chinese, Sri Lankan Sign Language. Having so many sign languages, most of the researchers has worked on single sign language. The face expressions and the hand movements can vary in sign languages. For example, British sign language uses both hands for gestures, where as American sign language uses single hand. This section will cover the past research and their limitations. The background and the methods used

for the research and how it can benefit the society.
The literature review is divided in three sub-sections, Section 2.1 will discuss the Sign language recognition and translation based on sensors and images. This section will also discuss various languages and the method of the processing. There are some studies which used hardware and sensors this will also be covered in this section. Section 2.2 will cover previous research and evolution of text-to-speech and Section 2.3 will go over the limitations of the literature on sign language recognition and text-to-speech technology that has been studied.

2.1 Literature on Sensor and Vision Based Methods

The recognition of sign language can be based on two methods one being the sensor based and the other is vison based (Mahmood and Abdulazeez; 2019). The approach based on the sensors requires the person to wear the sensor and then perform the signs whereas the vision-based approach uses the camera captured images which subsidises the need of carrying the sensors or gloves (Rautaray and Agrawal; 2015). Aly et al. (2019) stated that recently the use of the cameras at low cost enables the users to apply them in many researches and the use of image processing, machine and deep learning is exploring these fields to reduce the cost and expose these for greater good.

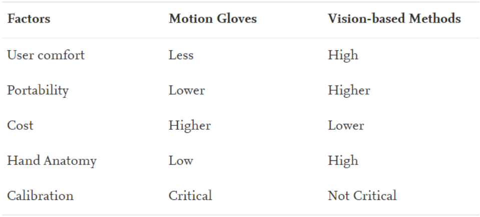
Sign language is the primary language of people who have hearing loss and difficulty speaking. Camgoz et al. (2018) extracted German Sign Language Translation from videos of sign language and converted to spoken language. In this case, the dataset includes a German sign language video of a weather broadcast as well as German sign vocabulary. Convolution neural networks and attention-based encoder-decoders are used to train the sign language. For better translation, word embedding is required. In comparison to other translation models, this model necessitates a large amount of training data. The output is average, and the accuracy is 44%, with some errors visible in the ground truth table.

A Times Series Neural Network-based real-time sign language translation was proposed by Kumar et al. (2018). This model is entirely based on OpenCV and a Neural Network. The goal is to convert video of sign language into English sentences. Continuous sign language video recognizes ASL gloss. The person must perform Sign Language in front of a webcam while wearing dark clothing and against a dark background. In a later stage, the hand and face are extracted from the frame; the hand contains a lot of information. A time series Recurrent Neural Network (RNN) is used. It can only operate on one frame at a time. For gloss-to-speech translation, encoders and decoders are used. Overall, this approach was successful in offering a new perspective on translation.

For German Sign Language translation, a support Vector Machine and a 3D object detection technique are used. Three network architectures are used in 3D object detection: localisation, elevation, and segmentation. This model teaches thirty hand gestures. The hand gesture dataset is labeled with 3D points, which aid in learning the pose. To recognize the hand gesture, Mohanty et al. (2020) used a three-network pipeline. PoseNet, HandSegNetand, and PosePrior are three network pipelines. There were no depth cameras used in the 3D estimation. Sign language is recognized after obtaining a hand pose representation. The Support Vector Machine in conjunction with the Radial Basis Function is the classification model used here. The Support Vector Machine model and 3D object detection were tested and found to be successful. The Support Vector Machine classifier alone had an error rate of 0.58. By combining features, the rate was reduced to

0.39. Real time american language recognition was built using CNN by Shahriar et al. (2018). Image processing is done using skin detection using YCbCr and RGB and the background is cropped is used. Multi layer perceptron layer is implemented in CNN. The accuracy achieved by the model is 94.7%. But if there is any object in background, this model has difficulty in detecting the signs.

In Lee et al. (2021), the author has compared the vision based and hardware based methods on various factors that needs to be considered while building the system. The below Figure 2 shows the comparison table. Based on the table we can conclude that the vision-based methods are more feasible to build and use.



Factors	Motion Gloves	Vision-based Methods
User comfort	Less	High
Portability	Lower	Higher
Cost	Higher	Lower
Hand Anatomy	Low	High
Calibration	Critical	Not Critical

Figure 2: Glove-based Vs Vision-based methods

In 2019, Muthu Mariappan and Gomathi used Indian sign language dataset for real time sign recognition using Region of Interest(ROI) and skin segmentation from OpenCV and Fuzzy c-means(FCM) clustering algorithm. The system was able to predict 40 words and mostly recognised static gestures with 75% accuracy. The computation time for FCM was high. Further use of CNN or RNN was suggested. Hameed et al. (2022) the author proposed model with radar and deep learning. The study was conducted on the British Sign Language dataset having 15 different signs. The extraction of 2D spatiotemporal features from images and then Deep learning model was applied to classify these images. The accuracy achieves was 90.07% with VGGNet model.

RAJ and JASUJA (2018) and Quinn and Olszewska (2019) built British Sign Language Recognition using Histogram Oriented Gradients(HOG). ANN and SVM was implemented respectively with the HOG and the accuracy achieved was exceptional. Deep learning technique has become the primary strategy to provide solution to various gesture recognition and computer vison problems(Pisharady and Saerbeck; 2015) (Li et al.; 2015). In recent years the use of different deep learning algorithms has increased for recognition of hand gesture, also due to the expressiveness, blockages and the speed of variation of hand movement the author Oyedotun and Khashman (2017) states that this field stays active for more improvements.

2.2 Literature on Text-to-Speech

For conversation to take place, speech or voice is the most promising medium. Assisting uneducated, people having difficulty with vision and speaking are the uses of the TTS technology. With the recent advancements, the Text-to-Speech(TTS) has gained alot of exposure. From detecting the text, converting it to speech and then the speech synthesis. Many researches has been conducted to study the speed, generation style, expressions of the speech. With the text to speech, the linguistic translation is also taking place. Tan et al. (2021) states the field is ever changing with expressive TTS, fast TTS, adaptive TTS, low resource TSS etc..

Zen et al. (2013) built speech synthesis using deep neural network and decision tree. The author stated limitations like fragmentation, expression was not efficient, there was context dependency when used the decision tree model. The performance of deep neural network was better but it was time consuming and complex to compute.As RNN requires additional parameters and CNN is much faster than RNN, Tachibana et al. (2018) built TTS with CNN model. The model took 15 hours to get trained and parameter tuning was suggested to reduce the time required to operate the system. Capes et al. (2017) conducted an experiment in which they utilised Apple's Siri and implemented speech synthesis using deep learning to reduce the latency, building voice for multiple languages and synthesis quality using optimisation.

Kumar et al. (2022) built a system to help people having difficulty with vision, using the idea of Optical Character Recognition , Raspberry Pi interpreter and voice synthesizer. The Raspberry Pi produced the voice through speaker. The model was tested on various languages as well and was success, but the model was bulky because of the infrastructure.

2.3 Limitations of previous research

Majorly studies has been carried on single sign language. The glove-based technology has been expensive and complex. It also requires the person to wear and carry the gloves which can be uncomfortable and restrictive sometimes. Some studies have been carried out on small and limited dataset. Most of the study has implemented single sign language recognition. The text to speech translation part has not been implemented which if implemeted can build effective communicating system.. Considering the above this study can uniquely contribute to the society in communicating with the specially abled people effectively.

3 Methodology

Sign language is used by people who face difficulty in speaking and hearing. To communicate use of hand movements and gestures are required. The most popular sign language is American Sign Language, but there are other sign language as well like Indian Sign Language, British Sign Language, Australian Sign Language and apart from these there are 300 more sign languages. As everyone is not able to understand sign language, there is need to interpret these sign language. Also, as there are different types of sign languages, there should be a system where all sign languages can be recognised and translated. This research tends to assist in recognising the American Sign Language (ASL) and Indian Sign Language (ISL) so that people having difficulty in speaking and hearing can communicate effectively. Recognition model for both sign language was developed using Convolution Neural Network (CNN). For text to speech translation a library from python was used which generates the audio from the system. Knowledge Discovery Database (KDD) methodology was followed, Figure 3 represents the steps taken for this research.

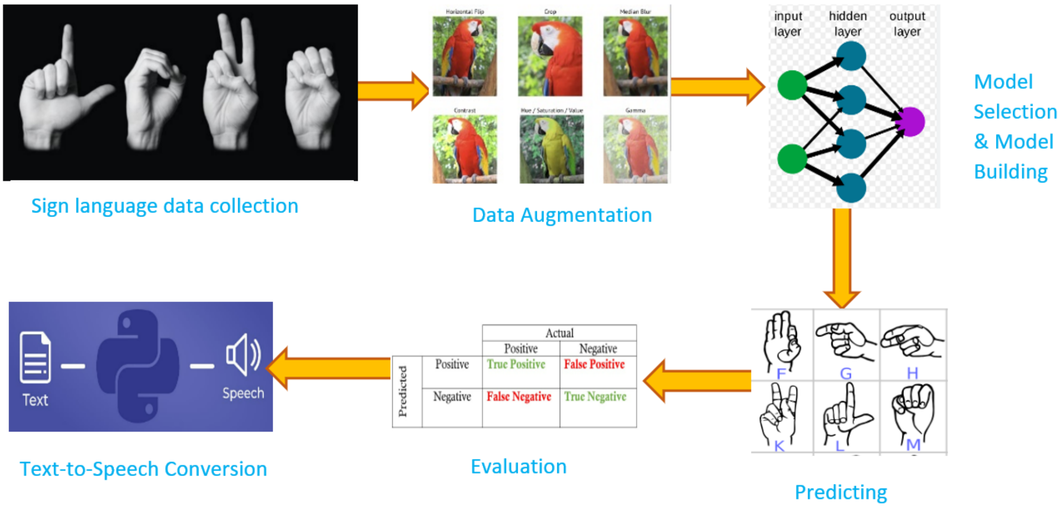


Figure 3: KDD for Sign Language Recognition

3.1 Data Selection

For this study, two dataset was collected one for ASL and one for ISL. The dataset was obtained from Kaggle. Both dataset has collection of images. The ASL dataset [5] has 29 folder each folder has 3,000 images of the A-Z characters and 3 extra signs are added, these are Space, Delete and Nothing. The dimension of each image is 200x200 pixels. The other dataset is the ISL dataset[6] which contains 35 folder and each folder has 1,200 images of A-Z characters and 1-9 numbers. The dimension of these images are 128x128 pixels. These dataset is split in 90:10 ratio for train and test. 90% of the data is used for training and 10% is used for testing.

3.2 Data Pre-processing

After selection of data, the necessary libraries were imported and the data was loaded into Jupyter notebook. The libraries imported were tensorflow, keras, cv2, numpy and matplotlib. There were some images which were aligned to extreme left and extreme right. For processing images the ImageDataGenerator package is imported. The transformation of images is called image augmentation. The transformation are performed on the original images and results in multiple transformed copies which are different from each other depending on the function performed on it. Each image is unique and the model is trained on various different images. Because of this, the overfitting issue is resolved and model is generalised.

4 Design Specification

This section will briefly describe the model used for training the data after image augmentation and end to end process of the system that is built. Convolution Neural Network was used for building the model and training the data.

[5] https://www.kaggle.com/datasets/grassknoted/asl-alphabet

[6] https://www.kaggle.com/datasets/vaishnaviasonawane/indian-sign-language-dataset

4.1 Modelling

With advancing Machine Learning and Deep Learning technologies, Convolution Neural Network is one of the most used architecture for image classification or detection. This is because CNN can reduce the parameters without damaging the quality of images. CNN can be built using multiple layers depending on the complexity and later can be used for dimensionality reduction. Figure 4 shows the basic architecture of CNN. As it is multilayer network the Conv layer can be reduced or increased according to the complexity of the image. The conv layer in CNN separates the features in the input image, the features seperated are used by the fully connected layer to identify the class of the image. CNN can consist of many layers which can include Convolution, Flatten, Dropout, Pooling, Normalisation and Dense. Apart from these there are activation functions like 'ReLu' and 'softmax'. The softmax activation function is used for output labels. Flattening is used to convert the 2D array in single-dimensional vector. Normalisation function is used for more independent learning. Dropout is used to deal with overfitting. Pooling is used for downsampling the image so that the image can be translation invariant. The dense layer is also called the fully connected layer. ReLU activation function is used as it is faster and non linear. Using these layers two CNN models were formed for ASL and ISL recognition. The image proccessed in the image aumentation is fed to the convolution model.

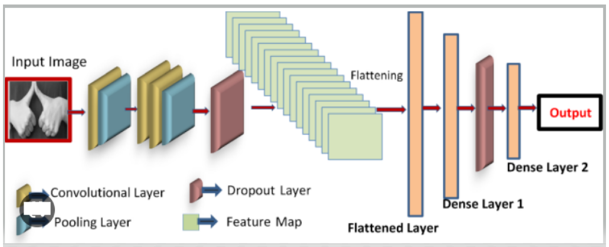


Figure 4: Basic architecture of CNN

4.1.1 CNN for ASL

For ASL, 5 layer CNN was implemented. The processed image is taken as an input and two conv layers are applied which extracts the features and pass it to the MaxPooling layer which is used to reduce the overfitting problem. The filters applied in Conv layer are 64 with kernel size as 5. The input shape is 64,64,1 and the output recieved after MaxPooling layer is 16,16,64. After this dropout layer with 0.5% is applied to drop the nodes. Again two conv with 128 filters, one MaxPooling and one dropout layer is applied. The output obtained is 4,4,128. The fifth conv layer has 256 filters followed by dropout and flatten function and output shape is 4,4,256 The dense layer is applied with 29 units as there are 29 classes. The padding value was defined to be 'same' The ReLU activation function and Adam optimiser were applied. The accuracy metrics was used and 3 epochs were executed with batch size 64. The above built CNN model was used with ASL model.

4.1.2 CNN fo ISL

10 layer CNN have been configured for ISL. The augmented image were passed to the convolution layer and feature extraction took place. two conv layer with 64 filters, two

with 128, three with 256 and three with 512 filters were built with kernel size 3x3, each layer has ReLU activation function and no padding was applied. Each layer was followed by MaxPool 3x3, the output shape obtained after these conv layer was 1,1,512. After this batch normalisation, flatten and dropout layers were applied. Normalisation is used for faster and independent learning. These layers were followed by 3 dense layer each having value of 1024, 512 and 36 respectively. The softmax function was applied for the 36 class output. The model used adam optimizer and was evaluated on accuracy metrics. Batch size was 32 and 3 epochs were executed.

4.2 Process Flow

Figure 5 represents flowchart of sign language recognition and translation. The steps for carrying out the experiment and obtaining the results are outlined below. A web app is created so that the experiment can be performed in a user-friendly environment.

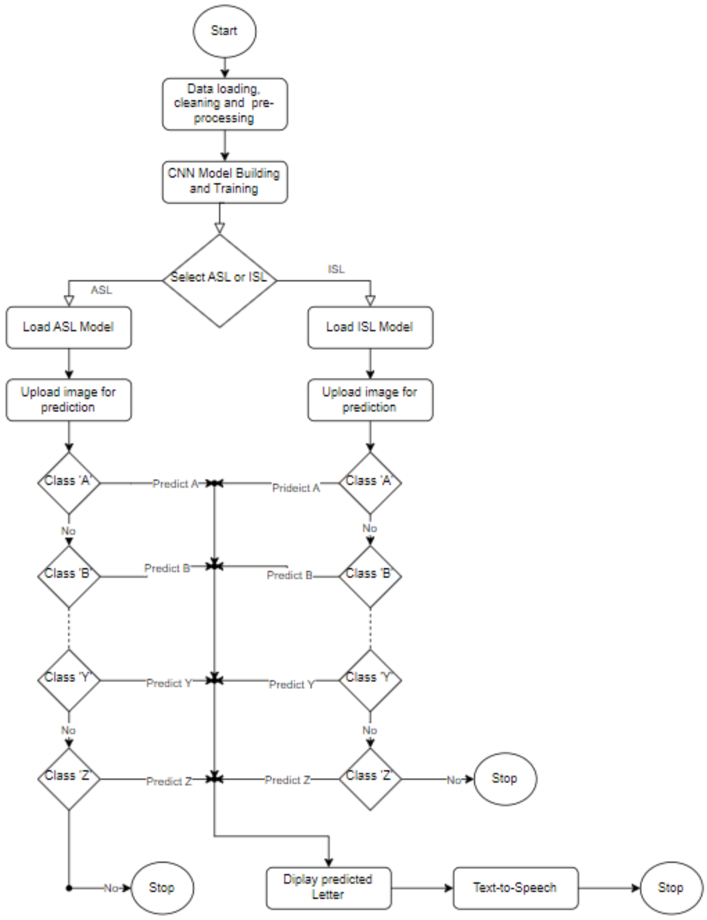
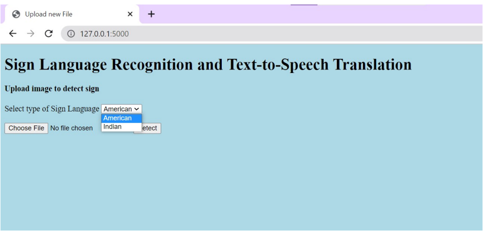


Figure 5: Flowchart for Sign Language Recognition

The web app displays a message asking user to select the language that he wishes to recognize. There is only one drop-down menu to choose between ASL and ISL. The prediction models are built using multi-layer CNN discussed in Section 4.1. The prediction model is run in accordance with the user language selection. After selecting a sign language, it will require you to upload an image of the sign for recognition.


Sign Language Recognition and Text-to-Speech Translation

Upload image to detect sign

Select type of Sign Language: American

Figure 6: Sign Language Recognition

Figure 6 shows the first page of the web app where it asks to select the language and upload image for recognition. Further steps are covered in Section 5.2 and Section 5.3

5 Implementation

This section will go over the experiments that were carried out and the results that were obtained. It will also cover the experiment setup, tools, and language used for this research. Section 3.1 and 3.2 has discussed the selection and processing of data. Section 4.1 has discussed the model applied on ASL and ISL. Section 4.2 represents and provides brief description of the overall process. This system is proposed for the recognition of Indian and American Signs. The experiments performed are described below.

5.1 Setup

Convolution Neural Network when performed on images takes a long time to process. Python was used to process the data and build the model. Keras, cv2 and Numpy is used for processing and augmenting the images. gtts (Google Text-to-Speech) library was used to convert the recognised characters using system generated audio. Jupyter Notebook was used to from Anaconda Navigator to load and perform these experiments. A system with 500GB hard drive and 8 GB RAM was used to execute the project.

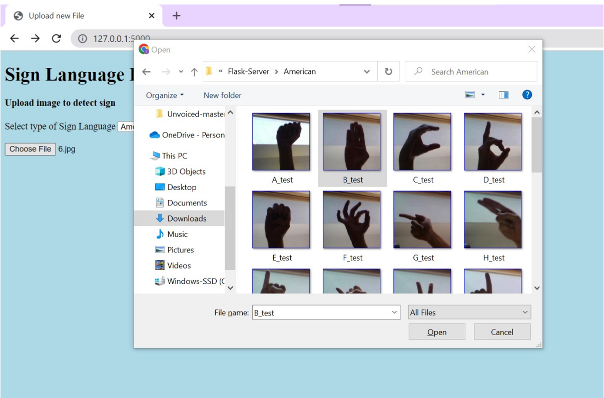
5.2 Experiment 1: American Sign Language Recognition and Translation

The first experiment was conducted with American Sign Language. A test image was selected and uploaded for recognition. Figure 7 displays the selection of an ASL image for recognition.

After the image has been loaded, there is a button that needs to be clicked for recognition. When the button is clicked, the predicted character is displayed and spoken out loud by the system itself. The detect button click triggers the prediction function for American Sign Language Figure 8 shows the successful prediction of an ASL image.

5.3 Experiment 2: Indian Sign Language Recognition and Translation

The second experiment was carried out using Indian Sign Language. For recognition, a test image was chosen and uploaded. The selection of an ISL image for recognition is



Sign Language I

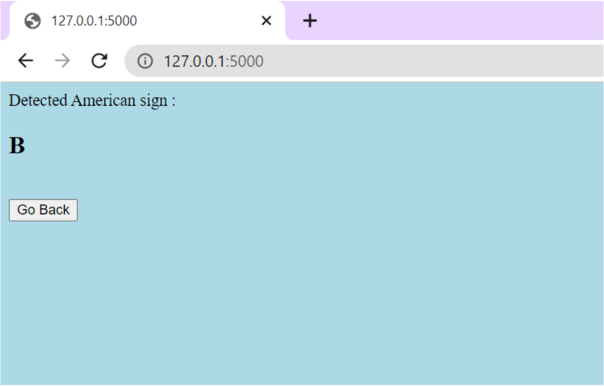
Upload image to detect sign

Select type of Sign Language Ame

Choose File 6.jpg

File name: B_test

Figure 7: ASL image upload

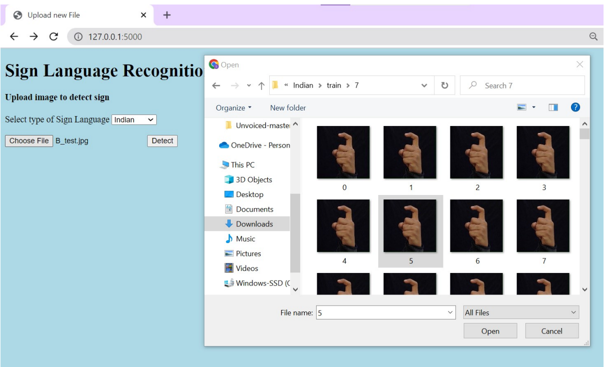


Detected American sign :

B

Figure 8: ASL prediction

displayed in Figure 9. The detect button triggers the prediction function of Indian Sign Language.



Sign Language Recognitio

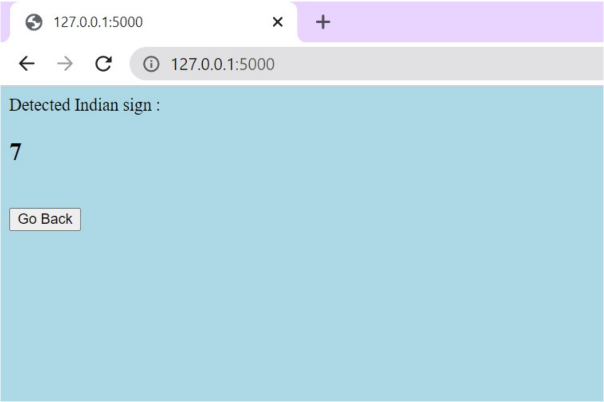
Upload image to detect sign

Select type of Sign Language: Indian

File name: 5

Figure 9: ISL image upload

After loading the image, 'Detect' button should be clicked for recognition. When you press the "Detect" button, the prediction function for Indian Sign Language is activated. The system displays the predicted character and speaks it aloud. Figure 10 shows the character prediction of ISL



Detected Indian sign :

7

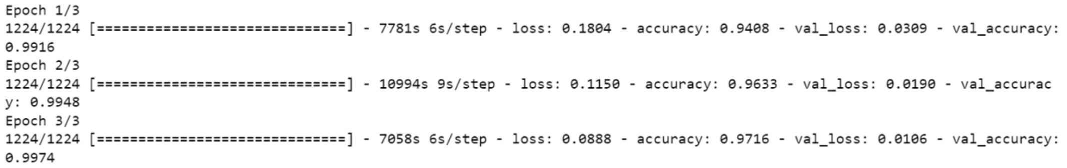
Figure 10: ISL prediction

6 Evaluation

This section covers the results obtained from the experiments performed in Section 5.2 and Section 5.3. The accuracy and losses are assesed at each epoch. The plots of accuracy and loss are displayed in Section 6.1 and Section 6.2 for ASL and ISL respectively.

6.1 Finding 1: American Sign Language Recognition and Translation

The images in the American Sign Language dataset were sent to the model for training and testing. The dataset was split into 90:10 ratio where 90% of the data was used for training and 10% was used for testing . The Figure 11 shows the accuracy of the model.



```
Epoch 1/3
1224/1224 [==============================] - 7781s 6s/step - loss: 0.1804 - accuracy: 0.9408 - val_loss: 0.0309 - val_accuracy:
0.9916
Epoch 2/3
1224/1224 [==============================] - 10994s 9s/step - loss: 0.1150 - accuracy: 0.9633 - val_loss: 0.0190 - val_accurac
y: 0.9948
Epoch 3/3
1224/1224 [==============================] - 7058s 6s/step - loss: 0.0888 - accuracy: 0.9716 - val_loss: 0.0106 - val_accuracy:
0.9974
```

Figure 11: ASL Epoch Accuracy

It can be seen that the accuracy and validation accuracy is increasing with every epoch and the loss is decreasing. Figure 12 represents the graph plotted for accuracy vs validation accuracy and training loss vs validation loss.

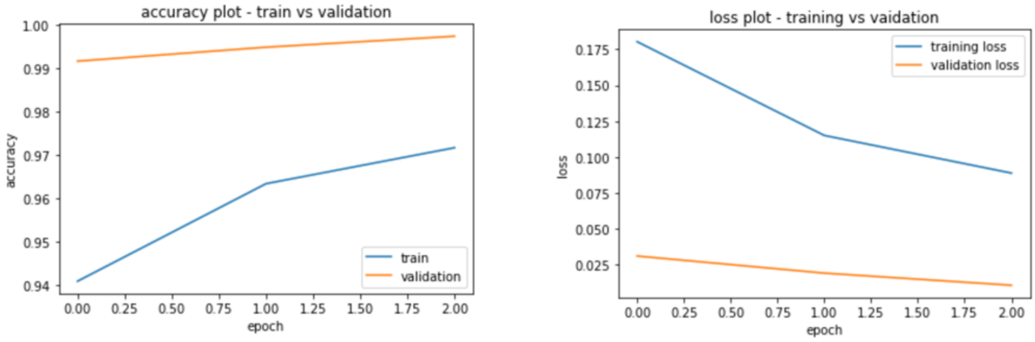
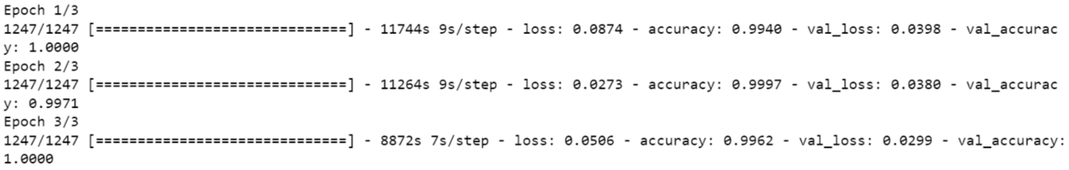


Figure 12: ASL Accuracy and Loss plot

From the loss plot it can be interpreted that the model has undefitting and is capable to learn more as the it continues to decrease till the end.

6.2 Finding 2: Indian Sign Language Recognition and Translation

The dataset had 35 folder and the images from the folders were used to train the model. The dataset was segmented in 90% training and 10% testing images.The Figure 13 shows the accuracy of the model.



```
Epoch 1/3
1247/1247 [==============================] - 11744s 9s/step - loss: 0.0874 - accuracy: 0.9940 - val_loss: 0.0398 - val_accurac
y: 1.0000
Epoch 2/3
1247/1247 [==============================] - 11264s 9s/step - loss: 0.0273 - accuracy: 0.9997 - val_loss: 0.0380 - val_accurac
y: 0.9971
Epoch 3/3
1247/1247 [==============================] - 8872s 7s/step - loss: 0.0506 - accuracy: 0.9962 - val_loss: 0.0299 - val_accuracy:
1.0000
```

Figure 13: ISL Epoch Accuracy

Figure 13 displays the accuracy acheieved at every epoch that was executed. It shows that at second epoch the accuracy reached 99.9% and at third epoch it went down at 99.6%

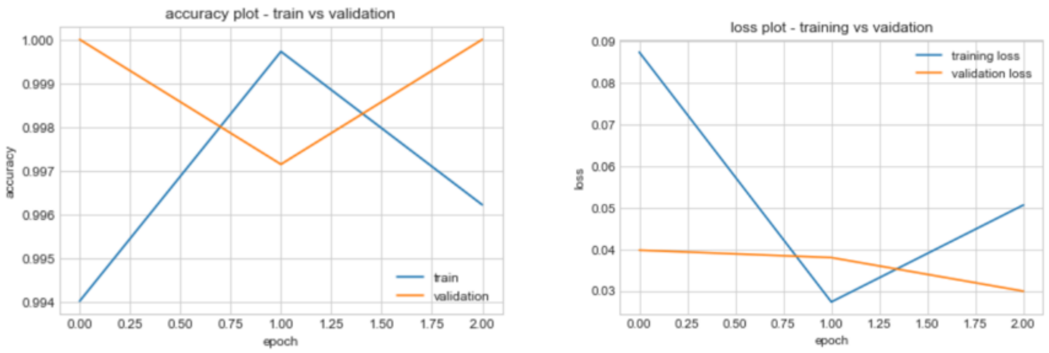


Figure 14: ISL Accuracy and Loss plot

Figure 14 shows the accuracy and loss graphs of the CNN model built for ISL Because there are only three epochs, the graph has sharp fluctuations. As accuracy improved, it was decided to limit the number of epochs at three.

6.3 Discussion

This research used one of the most used deep learning model for recognition of the signs. Convolutional Neural Network was implemented on ASL and ISL. Two different CNN architecture were built for image classification. This project was built to assist the people who have difficulty in speaking and listening. As for communication these two senses are crucial, conversing become difficult. This system is software based so there is no need of carrying glove or hardware. The system is web app based, the user needs to select the sign language he wants to recognise (ASL or ISL) and start uploading the pictures. After the upload when the user press detect, the sign in the image is recognised and spoken out loud by the system itself. There is no need to carry a speaker or any hardware for speech. This proposed system can help the normal people to decode the signs and understand what the specially abled person wants to convey.

This study has proposed a solution which is 99% accurate for ISL and 97% accurate for ASL. The accuracy and loss plot curves are discussed in the previous section. The prominent feature of this model is that it gives output in seconds and the prediction are accurate. Python library GTTS was used for text to audio conversion. So there is no extra cost or hardware required. The implementation was done on software and the computation time is less. The limitation of this study is that it is tested on image dataset. The recognition is performed on the images which in real-time can be frustrating. Also, the predictions are performed character by character. Also, manual intervention is required at selection and upload of the image.

In Mahmood and Abdulazeez (2019) various models were compared and new feature extraction was implemented. The accuracy achieved was 98%. Only black and white images were processed which has numbers from 0-10 in Kurdish language. Muthu Mariappan and Gomathi (2019) the author used ISL in realtime and FCM was implemented the computational time was high. The accuracy achieved was 75%. The data dictionary was small. There are many previous works which used sensors and gloves for recognition and prediction Rastogi et al. (2015) Aly et al. (2019)

7 Conclusion and Future Work

The motive of this research study was to assist people who are unfamiliar with American and Indian sign language to communicate. This is implemented so that the special-abled people having difficulty in hearing and speaking can communicate effectively. The model for recognition of the sign language was built using Convolution Neural Network. Two different models were built for Indian Sign Language and American Sign Language. Before building the model image processing was performed. The promising thing about this project is that it is web and software based. No extra hardwares are required and there is no computational cost. As there is no hardware requirements and it is totally software-based, it does not have an extra cost. It takes seconds to recognise the signs and read out loud which sign is detected. The accuracy achieved in American Sign Language is 97% with 3 epochs and Indian Sign Language is 99% with 3 epochs. The limitation of this research is that it uses images to recognise the signs. It becomes difficult to create word and sentences as every character is detected and spoken out loud. After the image is uploaded and detect is pressed it recognise the character and speaks out loud. At a time only single character is recognised. Manual intervention is required while selecting language, selecting image and uploading it which is a drawback in now advancing technologies.

In future the model can be applied on real time video feeds and the accuracy and prediction can be tested. More Sign Languages can be added so that it can be used over world. The manual interventions can be reduced by taking real-time feeds for recognition. The dataset is of characters, other dataset of words can be added so that sentences can be formed. A dictionary and spell checker can be added so that the words and sentences can be formed and recognised.

Aknowledgement

I am grateful to get the opportunity of studying at National College of Ireland and to conduct Research Project. The project support and supervision by Prof. Athanasios Staikopoulos was commendable. I appreciate the support from my friends and family and would like to mention Sagar Gatla and Ghansham Ghate for guiding, supporting and motivating me.

References

Aly, W., Aly, S. and Almotairi, S. (2019). User-independent american sign language alphabet recognition based on depth image and pcanet features, *IEEE Access* **7**: 123138–123150.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H. and Bowden, R. (2018). Neural sign language translation, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793.

Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarkhani, A., Hu, Q., Huddleston, N., Hunt, M., Li, J., Neeracher, M. et al. (2017). Siri on-device deep learning-guided unit selection text-to-speech system., *Interspeech*, pp. 4011–4015.

Hameed, H., Usman, M., Tahir, A., Ahmad, K., Hussain, A., Imran, M. A. and Abbasi, Q. H. (2022). Recognizing british sign language using deep learning: A contactless and privacy-preserving approach, *IEEE Transactions on Computational Social Systems* pp. 1–9.

Kumar, A. A., Senthilvasudevan, B. and Farhan, H. U. (2022). Translation of multilingual text into speech for visually impaired person, *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, pp. 60–64.

Kumar, S., Wangyal, T., Saboo, V. and Srinath, R. (2018). Time series neural networks for real time sign language translation, pp. 243–248.

Lee, C., Ng, K. K., Chen, C.-H., Lau, H., Chung, S. and Tsoi, T. (2021). American sign language recognition and training method with recurrent neural network, *Expert Systems with Applications* **167**: 114403.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417420310745*

Li, S.-Z., Yu, B., Wu, W., Su, S.-Z. and Ji, R.-R. (2015). Feature learning based on sae–pca network for human gesture recognition in rgbd images, *Neurocomputing* **151**: 565–573.

Mahmood, M. R. and Abdulazeez, A. M. (2019). Different model for hand gesture recognition with a novel line feature extraction, *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pp. 52–57.

Mohanty, S., Prasad, S., Sinha, T. and Krupa, B. N. (2020). German sign language translation using 3d hand pose estimation and deep learning, *2020 IEEE REGION 10 CONFERENCE (TENCON)*, pp. 773–778.

Muthu Mariappan, H. and Gomathi, V. (2019). Real-time recognition of indian sign language, *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pp. 1–6.

Oyedotun, O. K. and Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition, *Neural Computing and Applications* **28**(12): 3941–3951.

Pisharady, P. K. and Saerbeck, M. (2015). Recent methods and databases in vision-based hand gesture recognition: A review, *Computer Vision and Image Understanding* **141**: 152–165.

Quinn, M. and Olszewska, J. (2019). British sign language recognition in the wild based on multi-class svm, *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 81–86.

RAJ, R. D. and JASUJA, A. (2018). British sign language recognition using hog, *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1–4.

Rastogi, R., Mittal, S. and Agarwal, S. (2015). A novel approach for communication among blind, deaf and dumb people, *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, pp. 605–610.

Rautaray, S. S. and Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey, *Artificial intelligence review* **43**(1): 1–54.

Shahriar, S., Siddiquee, A., Islam, T., Ghosh, A., Chakraborty, R., Khan, A. I., Shahnaz, C. and Fattah, S. A. (2018). 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*, IEEE, pp. 1168–1171.

Tachibana, H., Uenoyama, K. and Aihara, S. (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4784–4788.

Tan, X., Qin, T., Soong, F. and Liu, T.-Y. (2021). A Survey on Neural Speech Synthesis. arXiv:2106.15561 [cs, eess] version: 3.
URL: *http://arxiv.org/abs/2106.15561*

Zen, H., Senior, A. W. and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* pp. 7962–7966.