SIGN LANGUAGE RECOGNITION

Joshitha Gandra, Vaishnavi Patil, Prof. Gayathri Ananthanarayanan

Abstract - Conversing to a person with hearing disability is always a major challenge. Deaf and Mute people use sign language to communicate, hence normal people face problems in recognizing their language by signs made. Generally sign language consists of hand gestures and facial expressions. In this, hand gestures were considered to recognize the letters and words. We have used different deep learning models to predict the sign.

1. Introduction

The study by the World Health Organization (WHO) reports that more than 7% of the world 's population has hearing impairment. It is estimated that about 900 million people will experience hearing loss in 2050. According to WHO 2018 report in India, about 63 million people suffer from hearing impairment.Those suffering from speech and hearing loss find it impossible to communicate with normal and vice versa. Use of sign language is the only way to communicate. But a normal person cannot understand sign language, and a device is required to translate a sign into a common regional language. Sign languages are visual representation of hand gestures, finger movements, facial expression, body movement, etc. Different countries have their respective form sign gesture communication which results in non-uniformity, i.e. Indian Sign Language (ISL) in India, American Sign Language (ASL) in America etc. Sign language structure varies spatially and temporally. There are sensor based methods and vision based methods. In vision based gesture recognition technology, a camera reads the movements of the human body, typically hand movements and uses these gestures to interpret sign language. In sensor based methods, real- time hand and finger movements can be monitored using the Leap motion sensor. We have worked on **Fingerspelling and World Level Sign Vocabulary** using Vision based and Leap motion controller bases systems**.**

3. Literature Survey

- Chuan CH, Regina E, Guardino C (2014) American Sign Language recognition using leap motion sensor. In: 13th IEEE international conference on machine learning and applications (ICMLA), pp 541–544

Chuan et al. developed an American Sign Language recognition system using leap motion sensor. The system was classified using K-Nearest Neighbor and Support Vector machine and the accuracy of 72.78% and 79.83% was achieved respectively.

- “Deep Convolutional Neural Networks for Sign Language Recognition”

G.Anantha Rao, Guntur (DT)

Extraction of complex head and hand movements along with their constantly changing shapes for recognition of sign language is considered a difficult problem in computer vision

- K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition.

2D CNN are used to extract spatial features of input images while RNN are employed to capture the long term temporal dependencies among input video frames. VGG16 pretrained on ImageNet to extract spatial features and then feed the extracted features to a stacked GRU.

- J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. CVPR, 2017.

3D convolutional networks are used which are able to establish not only the holistic representation of each frame but also the temporal relationship between frames in a hierarchical fashion. Inflate 2D filters of the Inception network trained on ImageNet, thus obtaining well-initialized 3D filters.

- Recognizing American Sign Language Gestures from within Continuous Videos

One of the main challenges is that in actions in continuous videos, the temporal boundaries of a specific movement are not very clear. This paper detects their temporal locations from within continuous videos, by collecting an ASL dataset that has been annotated with the time-intervals for each ASL word.

2. Existing Approach

Identification of sign gesture is performed with either of the two methods.

- **Glove based method** whereby the signer wears a pair of data gloves during the capture of hand movements.
- **Vision based method**,further classified into static and dynamic recognition. Static deals with the detection of static gestures (2d-images) while dynamic is a real time live capture of the gestures.

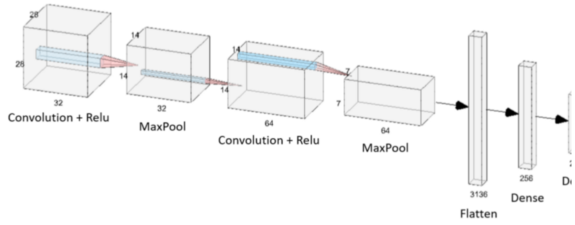
And despite having an accuracy of over 90%, wearing gloves are uncomfortable and cannot be utilised in rainy weather. They are not easily carried around.

4. Methodology

A. Static Hand gesture Recognition - fingerspelling

a. Convolutional Neural Network

A Convolutional Neural Network (CNN) is comprised of one or more convolutional layers and then followed by one or more fully connected layers as in a standard multilayer neural network. The architecture of a CNN is designed to take advantage of the 2D structure of an input image. This is achieved with local connections and tied weights followed by some form of pooling which results in translation invariant features. Another benefit of CNNs is that they are easier to train and have many fewer parameters than fully connected networks with the same number of hidden units. CNNs are very effective in reducing the number of parameters without losing on the quality of models. Images have high dimensionality which suits the above described abilities of CNNs.

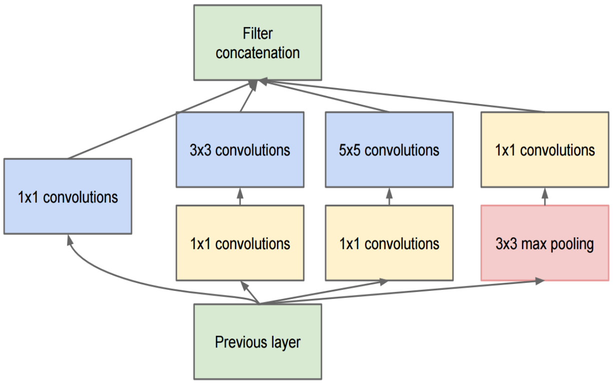


b. Transfer Learning: Inception V3 model

Transfer Learning : A pre-trained model that has been trained on an extremely large dataset is used, and we transfer the weights which were learned through hundreds of hours of training on multiple high powered GPUs.

We have used Inception -V3 Model for classification: Trained using a dataset of 1,000 classes from the original ImageNet dataset which was trained with over 1 million training images.

The main difference between the Inception models and regular CNNs are the **inception blocks**. These involve convolving the same input tensor with **multiple filters** and **concatenating** their results. On the contrary, regular CNNs performs **a single** convolution operation on each tensor.



LIMITATIONS OF STATIC DATASETS:

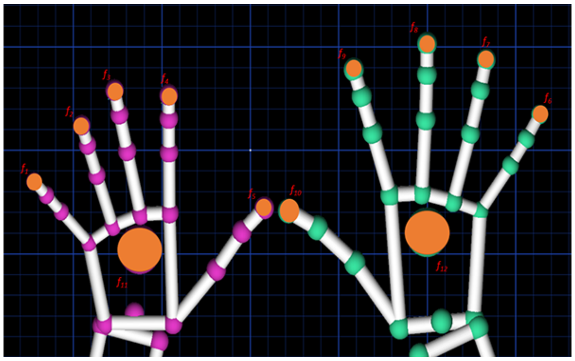
- Inability (or less accuracy) to recognize moving signs, such as the letters "J" and "Z".
- Fingerspelling for big words and sentences is not a feasible task.
- Temporal properties are not captured.

B. Dynamic Hand gesture Recognition - Leap Motion Controller

The leap motion technique makes use of a sensor, called Leap Motion Controller, based system. Here the information about hand and finger movements is captured by this sensor via APIs designed for the same. This is done by performing the movements a few feet above the horizontally positioned sensor. This data is then sent to a computer via USB.

Objective : To identify continuous single or double handed sign gestures using the Leap motion sensor.

The dataset has 42 different sign words, 90 fps. 12 dynamic features for each signer have been extracted that generate the sign gesture for both hands. Every feature is 3D coordinates of the respective fingertips.

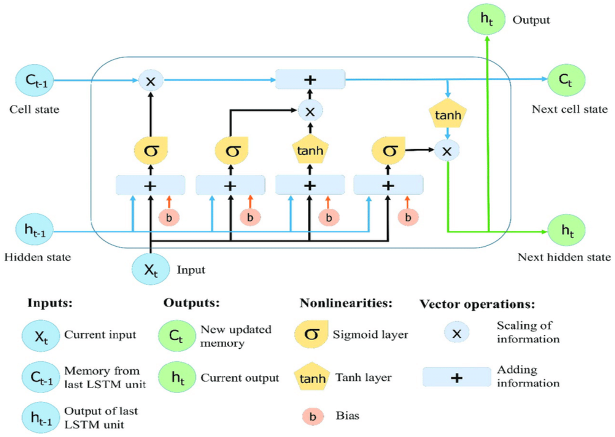


3D Coordinates extracted using Leap motion (orange color)

Hand gestures were considered to recognize the words. We have used two machine learning models: CNN and LSTM

Long Short Term Memory (LSTM) Model

The LSTMs are specifically designed to avoid the problem of long-term dependence. Remembering information is practically their default behavior for long periods of time.



Pictorial representation of LSTM model with various blocks

A number of sign language recognition (SLR) systems have been developed by researchers, but they are limited to isolated sign gestures only. Here we use a modified long short-term memory (LSTM) model for continuous sequences of gestures or continuous SLR that recognizes a sequence of connected gestures.

The model we used consists of two layers of LSTM with Rectified linear unit (ReLU) activation, followed by dropout and dense layers together with one Softmax layer for multi-class classification. Adam optimizer was used which was found to be robust against noisy class gradients.

CAMERA vs LEAP MOTION SENSOR BASED APPROACHES

Camera based approach is successful in detection of one hand, both hands and static or dynamic images (both isolated signs or continuous signs). It's easier to be adopted, cost effective. Leap Motion Sensor based is not restricted to 2D image/video capturing, but can also capture depth information such as color depth, etc. effectively. It's maintenance and overall costs pose a higher overhead than the camera method and is not feasible in real-world situations.

C. Dynamic Hand gesture Recognition -Video Based Approach

OBJECTIVE : Applying video classification on the video dataset of ASL signs. Take the captured videos, break down into frames of images that can then be passed onto the system for further analysis and interpretation.

We use the **WLASL 2000 dataset**.The videos have lengths ranging from 0.36 to 8.12 seconds, and the average length of all the videos is 2.41 seconds.

Why WLASL?

- Large Vocabulary size
- Other datasets fail to capture the difficulties of the task due to insufficient amount of instance and signer

I3D Model

We have employ the network architecture of I3D.

Transfer Learning : The original I3D network is trained on ImageNet and fine-tuned on Kinetics-400.

Weights and biases were taken from Inception v3 model(pre-trained from ImageNet Dataset). When inflating the 2D model into 3D, they simply took the weights and biases of each 2D layer and "stacked them up" to form a 3rd dimension.
The final I3D architecture was trained on the Kinetics dataset, a massive compilation

of YouTube URLs for over 400 human actions and over 400 video samples per action.

Data Preprocessing:

- Convert to mp4, extract Youtube frames and create video instances.
-Total number of videos for training: 18216, Total number of videos for testing: 2879
-Mode used: RGB Frames [Other option: optical flow]
(Dataset consists of RGB-only videos)

Model Set Up:

In order to better model the spatio-temporal information of the sign language, such as focusing on the hand shapes and orientations as well as arm movements, we need to fine-tune the pre-trained I3D.

Original model: Inception I3d (400, in_channels=3) : 400 classes and 3 input channels

Tuned Model : The class number varies in our WLASL dataset, only the last classification layer is modified in accordance with the class number.So, the last layer is replaced from 400 to 2000 neurons.

5. Results and Analysis

Dataset	CNN Layers	Accuracy
MNIST Dataset (28x28)	3 - layer CNN	Train : 95.3% Test : 94.7%
Alphabet Images (200x200 pixels)	3-layer CNN 5-layer CNN	Train: 99.06% Test: 98% Train: 97.29% Test: 99.816%

INFERENCES:

1) Two signs of letters such as“M” and “S” are confused and the CNN has some trouble distinguishing them.

2) Images of high resolutions are used in the Dataset 2 and hence the increase in the model accuracy is seen. As more pixels, allow for more intricate details to be extracted from the images..

Dataset	**Model**	**Accuracy**
ALPHABETS + (0-9) digits (200x200) 36 classes	5-layer CNN	Train: 97.1% Test: 98.1%
ALPHABETS + (0-9) digits + 15 static words 51 classes (200x200)	3-layer CNN	Train: 79.25% Test: 64.84%
ALPHABETS + (0-9) digits + 15 static words 51 classes (200x200)	Inception V3	Train: 98.99% Test : 94.88%

Inferences:

More object classes makes distinction between classes harder. Additionally, a neural network can only hold a limited amount of information, meaning if the number of classes becomes large there might just not be enough weights to cope with all classes. This justifies the reduction in model accuracy after adding more classes and training data in the dataset. The Inception V3 model has helped increase the model accuracy.

42 sign gestures (Some words and alphabets). Coordinates of gestures as acquired from leap motion sensor.

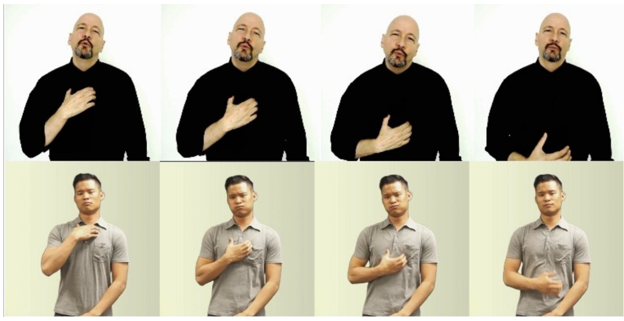
Algorithm	Accuracy
CNN (3-Layer)	**89.47%**
LSTM	**92.98%**

LSTM gives higher accuracy compared to CNN

For Video Classification:

Evaluation Metrics: Mean scores of top-K classification accuracy with K = {1, 5, 10} over all the sign instances.

As seen in Figures, different meanings have very similar sign gestures, and those gestures may cause errors in the classification results. However, some of the erroneous classification can be rectified by contextual information. Therefore, it is more reasonable to use top-K predicted labels for the word-level sign language recognition.



The verb “Wish” (top) and the adjective “hungry” (bottom) correspond to the same sign

Video classification results:

	Top-1	Top-5	Top-10
I3D Algorithm	**40.6%**	**71.58%**	**81.03%**

6. Conclusion

Here, we have used different deep learning and transfer learning models to predict sign letters and words. We tested the models on different datasets. In the future, we can use more signs. More signs can lead to more sentences and an improved model can be made to frame a correct sentence using minimum words. This models can be used in real life for hearing and speech impaired people to ease the communication between them and the world.