

RnD Project : Jan - April 2021

SIGN LANGUAGE DETECTION

Team :

- Vaishnavi Patil (180020039)
- Joshitha Gandra(180020009)

Supervisor:

Prof. Gayathri

PROBLEM STATEMENT



Conversing to a person with hearing disability is always a major challenge. Deaf and Mute people use hand gesture sign language to communicate, hence normal people face problem in recognizing their language by signs made. Hence there is a need of the systems which recognizes the different signs and conveys the information to the normal people.



EXISTING METHODS

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

- 1) **Glove based method** whereby the signer, wears a pair of data gloves during the capture of hand movements.
- 2) **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 of gloves are uncomfortable and cannot be utilised in rainy weather. They are not easily carried around since their use require computer as well.

Ref: [1]Sign Language Recognition using Sensor Gloves, FAST- National University of Computer and Emerging Sciences, Lahore



OUR METHODOLOGY

In this case, we have decided to go with the Vision based approach of hand gestures recognition, and improve on accuracy using Machine Learning and Deep Learning Algorithms. **American Sign Language(ASL) datasets** have been used for classification.

Sign language is a visual language and consists of 3 major components:

- 1) Fingerspelling : Spell out words character by character, and word level association which involves hand gestures that convey the word meaning. Static Image Dataset is used for this purpose.
- 2) World level sign vocabulary : The entire gesture of word or alphabets is recognised through video classification. (Dynamic Input / Video Classification)
- 3) Non-manual features: Facial expressions, tongue, mouth, body positions

We have worked on the first 2 components of **Fingerspelling and World Level Sign Vocabulary.**



STATIC HAND GESTURE DETECTION



STATIC HAND GESTURE DETECTION : FINGERSPELLING

Objective :

Producing a model which can recognise Fingerspelling based hand gestures in order to form a complete word by combining each gesture.

A. Alphabets (A-Z) in American Sign Language

Data Collection & Pre-Processing:

Dataset 1: **MNIST Dataset** : 28x28 pixels images(24 alphabets: J and Z deleted as they include gesture movements: ([Dataset](#)) [Training: 27,455 , Testing: 7172]

Dataset 2: **Image Dataset** : 200x200 pixels images: 29 classes, of which 26 are for the letters A-Z and 3 classes for *SPACE*, *DELETE*, and *NOTHING*. [Dataset](#) (J and Z were converted to static gestures by converting only there last frame)

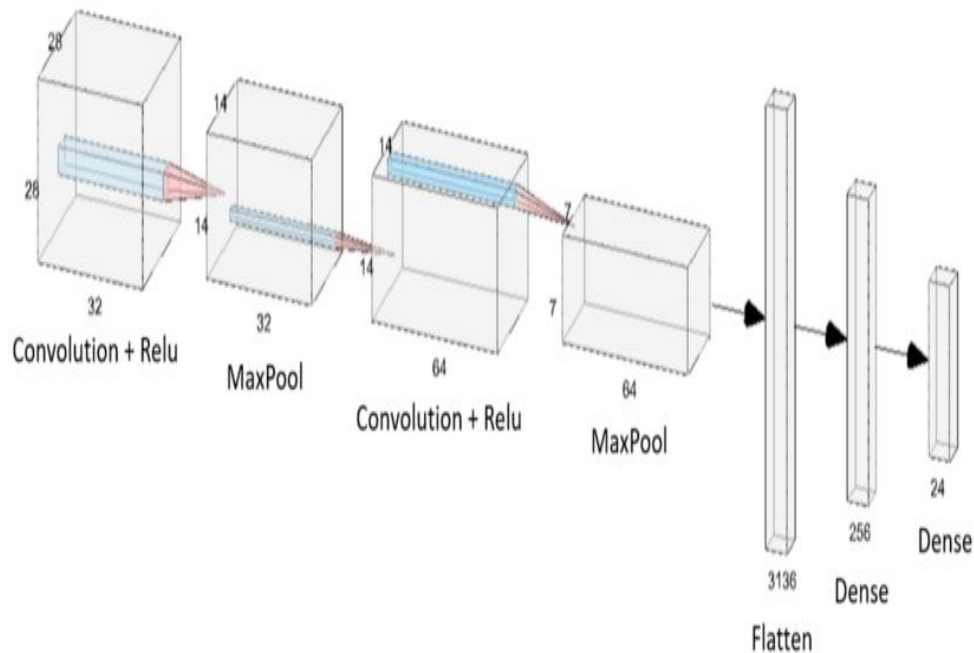
Learning/Modeling

We used a Convolutional Neural Network, or CNN, model to classify the static images in our first dataset.



WHY CNN?

- CNNs are very effective in reducing the number of parameters without losing on the quality of models. Images have high dimensionality (as each pixel is considered as a feature) which suits the above described abilities of CNNs.
- CNNs retains the 2D spatial form of images.
- All the layers of a CNN have multiple convolutional filters working and scanning the complete feature matrix and carry out the dimensionality reduction. This enables CNN to be a very apt and fit network for image classifications and processing.



RESULTS

Dataset	Training Set	Testing Set	CNN Layers	Accuracy
1) MNIST Dataset (28x28 pixels)	27,455	7172	3 - layer CNN	Train : 95.3% Test : 94.7%
2) Alphabet Images (200x200 pixels)	78,300	8700	A. 3-layer CNN B. 5-layer CNN	A.Train: 99.06% Test: 98% B. 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..

B. APPENDING THE DATASET CLASSES

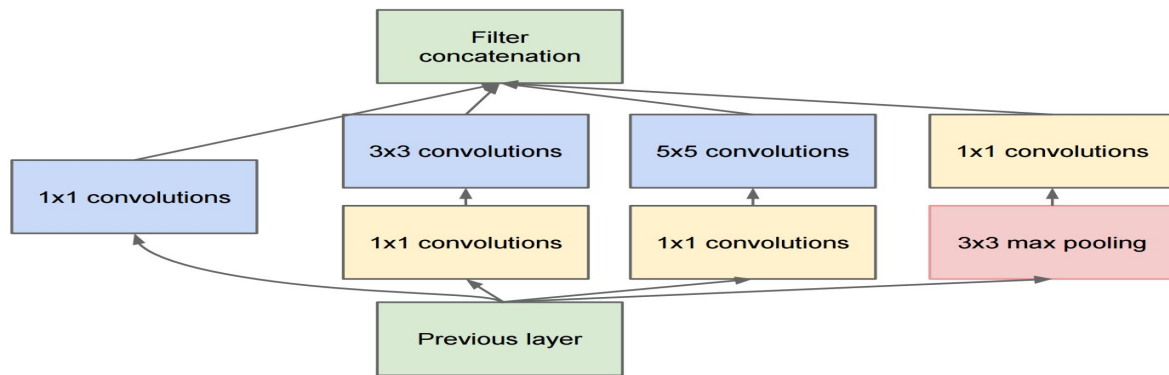
	Training Set	Testing Set	Model	Accuracy
ALPHABETS + 10 (0-9) digits (200x200 pixels Images) : 36 classes Dataset	24026	1265	A.3-layer CNN B.5-layer CNN	B.Train: 97.1% Test: 98.1%
ALPHABETS + 10 (0-9) digits + 15 static words (Baby, brother, etc.) : 51 classes (200x200 pixels Images) Dataset	182700	20300	3-layer CNN	Train: 79.25% Test: 64.84%

Inferences:

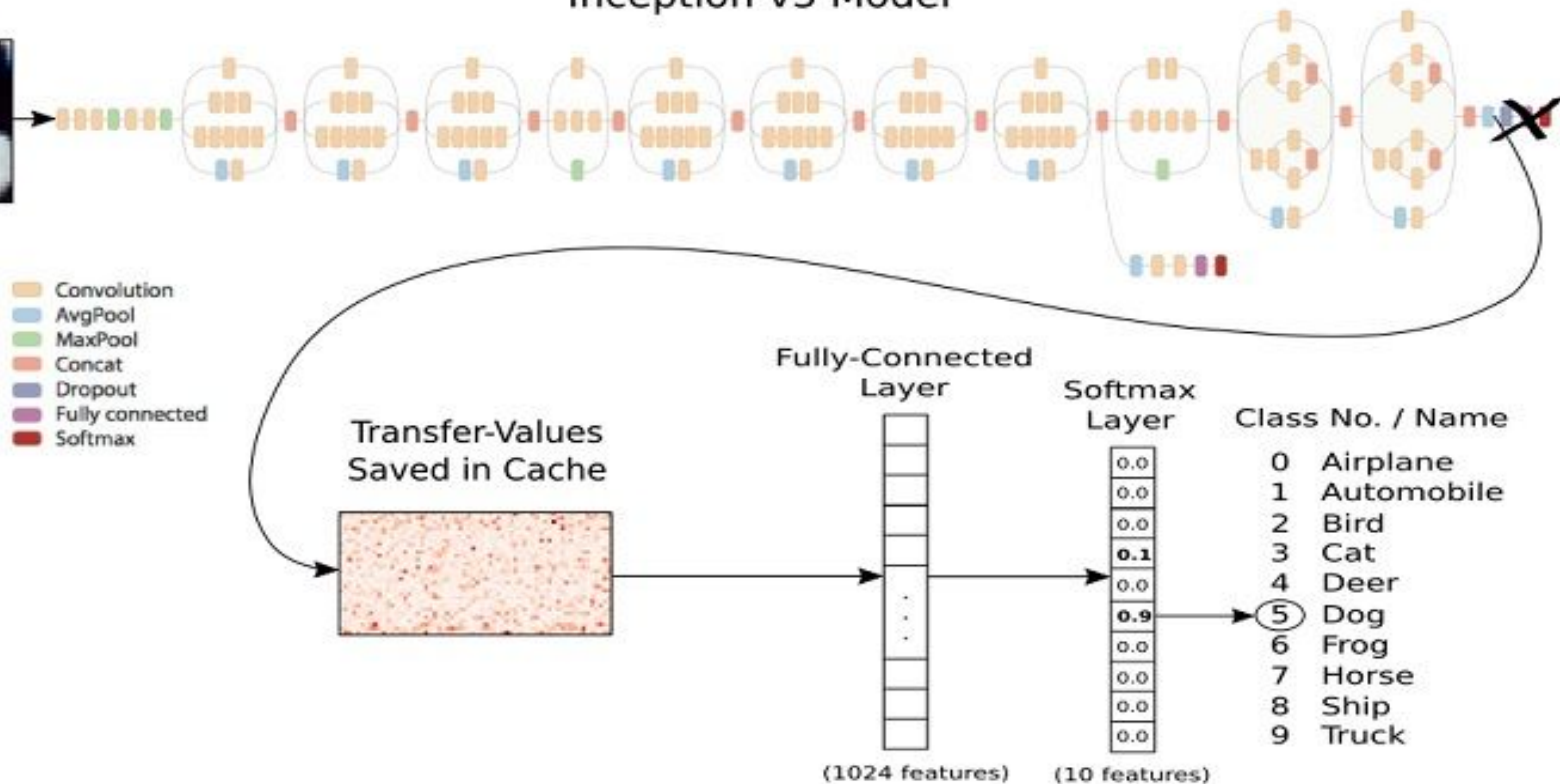
- More object classes makes distinction between classes harder. Additionally, a neural network can only hold a limited amount of information, meaning if the number of classes becomes large there might just not be enough weights to cope with all classes. This justifies the reduction in model accuracy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

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 contrast, regular CNNs performs **a single** convolution operation on each tensor.



Inception v3 Model



RESULTS

ALPHABETS + 10 (0-9) digits + 15 static words (Baby, brother, etc.) : 51 classes (200x200 pixels Images) Dataset	MODEL	ACCURACY
	3-layer CNN	Train: 79.25% Test: 64.84%
	Inception V3 Model	Train: 98.99% Test : 94.88%

Thus, the Inception V3 model has helped increase the model accuracy.



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.

The next phase of our project will focus on dynamic signs (i.e. moving signs).



Leap Motion Controller, based system

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.



DYNAMIC GESTURE DETECTION:LEAP MOTION

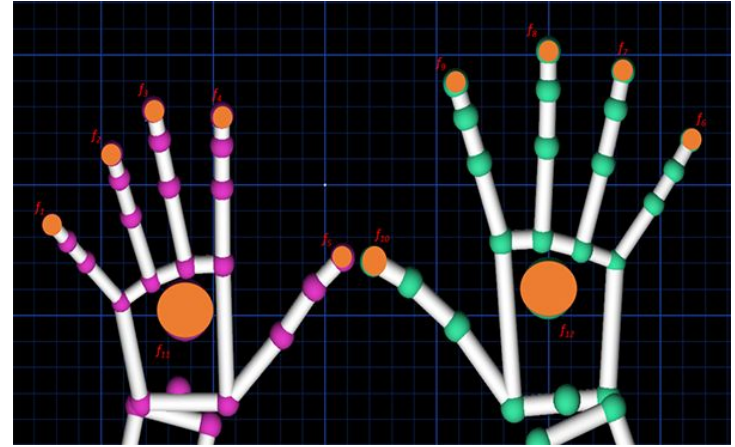
Objective : To identify continuous single or double handed sign gestures using the Leap motion sensor.

The dataset has 42 different sign words, 90 fps.

12 dynamic features for each signer have been extracted that generates the sign gesture for both hands. Every feature is 3D coordinates of the respective fingertips.

Ref: 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.

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



3D Coordinates extracted using Leap motion
(orange color)

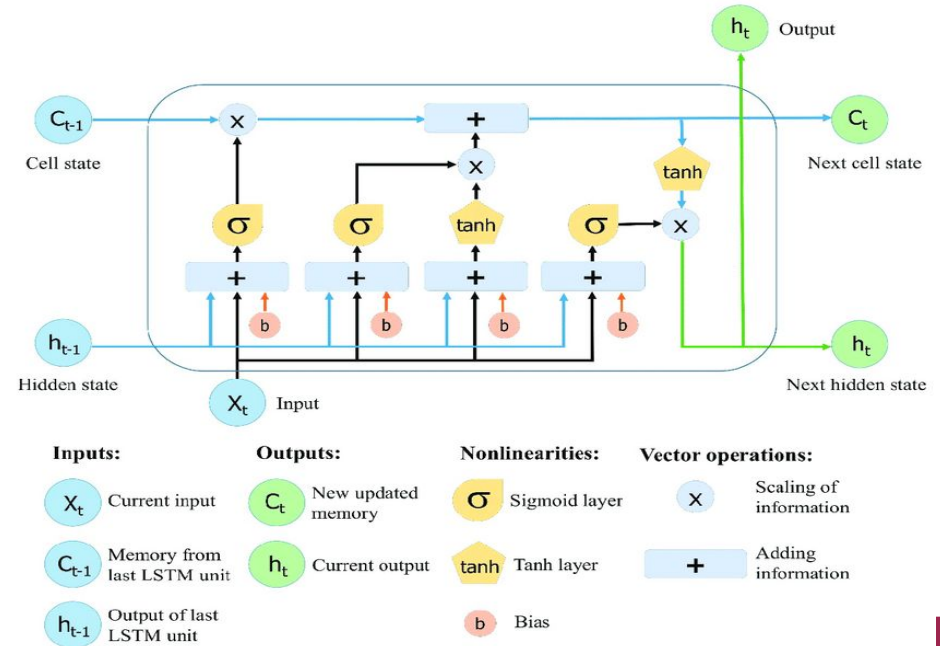
Hand gestures were considered to recognize the words. We have used two machine learning models:

- CNN and LSTM to predict the sign.

WHY LSTM?

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



Long Short Term Memory (LSTM) Model

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.



Results

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

Algorithm	Training Data	Testing Data	Accuracy
CNN (3-Layer)	264 (211: Training + 53 Validation)	114	89.47%
LSTM	264 (211: Training + 53 Validation)	114	92.98%

LSTM gives higher accuracy compared to CNN



CAMERA vs LEAP MOTION SENSOR BASED APPROACHES

Method ->	Camera or video capturing Based	Leap Motion Sensor Based
Advantages:	<ul style="list-style-type: none">- Successful in detection of one hand, both hands and static or dynamic images. The signs (both isolated signs or continuous signs) .- Easier to be adopted, cost effective.	<ul style="list-style-type: none">- It is not restricted to 2D image/video capturing, but can also capture depth information such as color depth, etc. effectively
Disadvantages:	<ul style="list-style-type: none">- Unable to capture depth information.	<ul style="list-style-type: none">- The maintenance and overall costs pose a higher overhead than the camera method.- Not feasible in real-world situations such as on roads, ships, shopping centres, etc- Unable to capture facial features and symbols which can be easily done in camera-based systems.

VIDEO BASED APPROACH : WL ASL DATASET

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.

Dataset Survey:

Comparison of available World-level datasets:

Dataset	#GLOSS	#Videos	#Signers
Purdue RVL-SLLL ASL Database	39	546	14
Boston ASLLVD	50	483	3
RWTH-BOSTON-400	2742	9794	6
WLASL 2000	2000	21,083	119

Why WLASL?

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

We propose to 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.

LITERATURE SURVEY

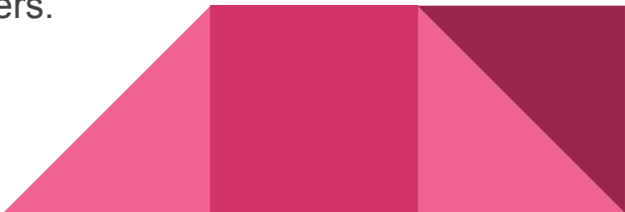
Signing, as a part of human actions, shares similarities with human action recognition. Some relevant works on action recognition:

[1] 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.

[2] 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.



OUR METHODOLOGY : I3D Model

We have employ the network architecture of I3D. [Ref : [Ref](#)]

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 // In the original kinetic dataset [Ref](#)

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.

RESULTS

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.

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



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

CONCLUSION

- Static Image Classification
 - Fingerspelling
 - Limitations and Importance of temporal features
- Dynamic
 - Leap motion
 - Video Classification



FUTURE WORK

In our future work, we also aim at utilizing word-level annotations to facilitate sentence-level and story-level machine sign translations



THANK YOU

