A Report on:-

Video based Gesture Recognition

Team Members :

1. Ashwin Vaswani (2017A7PS0960G)

Prepared in partial fulfillment of the Study Project under
Prof. Sujith Thomas:
Course No. CS F266

AT:-

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCES, PILANI
NOV 2019

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Contents:

INTRODUCTION :

With the recent advancements in the field of Machine Learning and Artificial Intelligence, there has been an upcoming trend of tools and robots to perform activities previously performed by humans. For this, the machine needs to understand the requirements and language of the user and what the user wants to communicate.

This was done in the past by giving a fixed set of instructions to the computer, but recent emphasis has been on making these machines smarter or “intelligent” to perform even more complex activities by adapting and learning as per requirement. This has led to a burst of research recently in the field of Human Computer Interaction.

Communication with machines using speech and written text has been widely explored in the past with already existing virtual assistants such as Siri and Alexa already a major part of our lifestyle now. However, there has not been concrete work on the use of actions or gestures to communicate with a machine to signal it to perform a set of activities. The following work aims at solving this problem and classifying hand gestures or actions to perform certain tasks.

COMPUTER VISION :

Computer Vision is a scientific field that deals with how computers understand and make inferences from visual data, that is, data in the form of images and videos. There has been a great boost in computer vision research over the recent years with the introduction of deep learning techniques to solve classical computer vision problems.

Computer Vision is not widely used to solve a number of problems such as Image Classification, Object Detection, Image Segmentation etc. These basic tasks are now serving as building blocks for further complex projects being carried out by organisations for automating our day-to-day activities and making

our life better. Along with its innumerable industrial use cases, there has been a lot of focus of computer vision in the field of robotics and incorporating computer vision techniques to make smarter and more intelligent humanoid robots in an attempt to create “Artificial humans” that can perform human activities with the same effect or even better than humans in some cases, with the extra potential that they possess.

GESTURE RECOGNITION :

Previous Work :

Before the introduction of convolutional neural networks and modern architectures for object detection, the task of gesture recognition was handled by detecting keypoints on the hand and then classifying gestures based on the orientation of those keypoints. The procedure involved segmenting the hand from the captured image and extracting the contour from it. However, this method had its limitations and performed poorly as the number of gestures increased and was not robust to variety in the types of background. Also, this method did not perform well for dynamic gestures which are generally required for progress in the domain of humanoid robots. This called for the need of deep learning algorithms and methods for effective gesture recognition.

Static Gesture Recognition :

Classical Static Gesture Recognition is performed by capturing the frame from the video stream and predicting the gesture on the basis of the loaded classifier model. A machine learning model needs to be trained beforehand which takes an image as input and predicts the class to which it belongs. This involves a number of different techniques such as convolution, pooling , normalisation, etc. Such a classifier however, is not feasible in real life situations because real life images contain a lot of noise in the background and thus some processing needs to be

done to handle this. For this purpose, our approach was to detect the hand and extract it from the frames. We then applied background subtraction followed by thresholding to get only the segmented hand from the image and then used these images to train our classifier. The resulting model gave outstanding results and was able to accurately classify 28/29 gestures it was trained on with an accuracy of more than 99%. To handle the problem of mispredictions or false predictions, we used a buffer window and took the max of the predictions in that buffer window which eliminated the issue of outlier predictions.

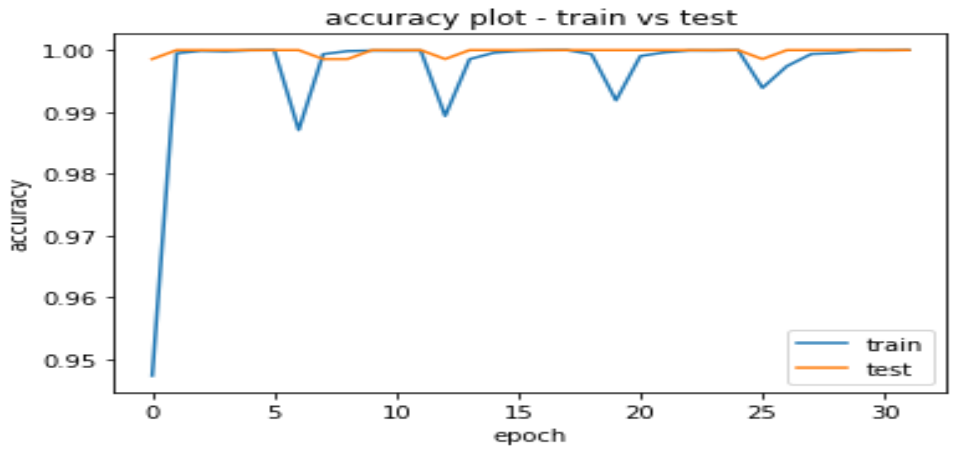
Dynamic Gesture Recognition :

As mentioned above, static gesture recognition required the use of 2D convolution operations to perform its classification. But, Dynamic gestures involve a chunk of frame and not just a single frame and these frames are dependent on each other over time and thus it is not right to use the same pipeline for dynamic gestures. Our approach was to use 3D CNNs for performing this task and 3D CNNs are virtually performing the same task as 2D CNNs but over an additional dimension, that is, it is now handling videos instead of images. We created a custom classifier model to detect and recognise these dynamic gesture movements and on further model optimisation, we were able to achieve an accuracy of around 85% on a 5 way classification task. Further improvements in model performance and number of classes were possible but limited by computational constraints. This model was then loaded on a live camera stream to perform classification and then these mapped gestures were used to perform various activities on the mobile device.

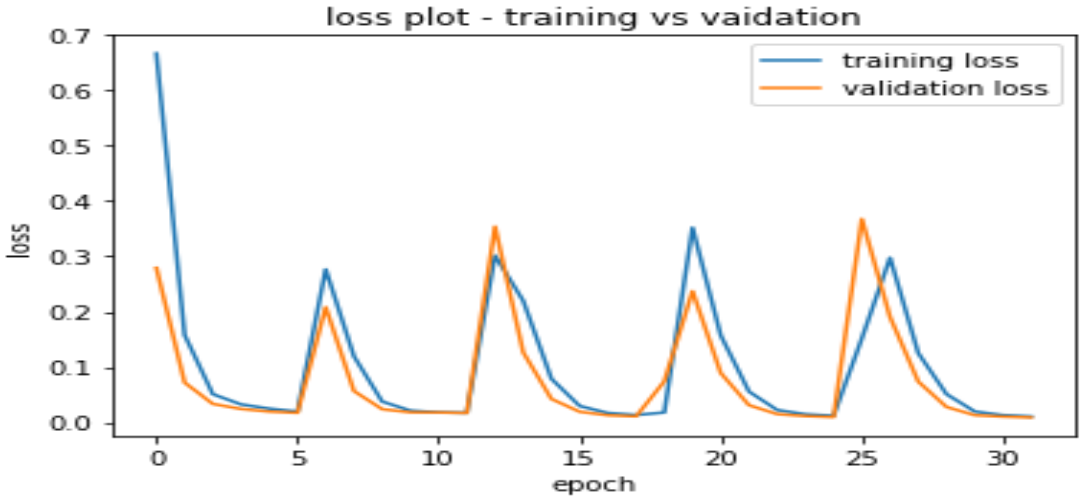
MODEL TRAINING :

SSD using Tensorflow’s Object Detection API was used to detect the hand from the images.
This extracted hand was then processed and finally passed to a custom classifier prepared by us. The model was trained to predict 29 classes and was run for 200 epochs using Adam as optimiser. The model finally achieved an accuracy of over

99% on the validation set. The loss and accuracy plots for the same are as follows:



Accuracy plot (Static)

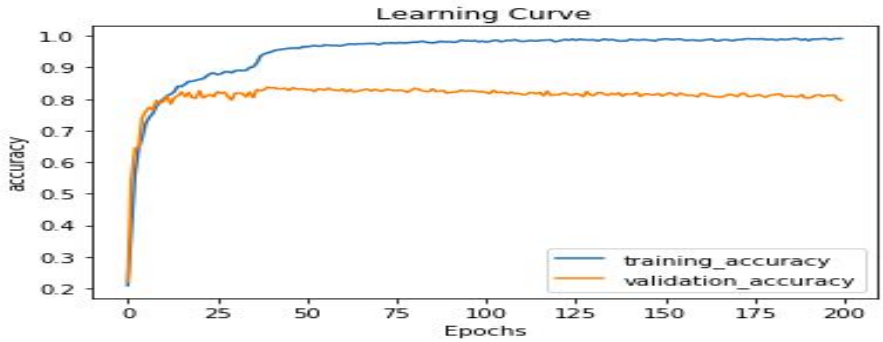


Loss plot (Static)

For training the model for Dynamic Gestures, we first searched upon available datasets and then decided to proceed with the 20bn Jesture dataset for our problem. We processes the images and selected 5 gestures for classification due to computational constraints and selected a chunk size of 16 frames for a video. We then trained the model using a custom made 3D-CNN architecture and saved the model. Finally, we loaded the model for making predictions on a live video

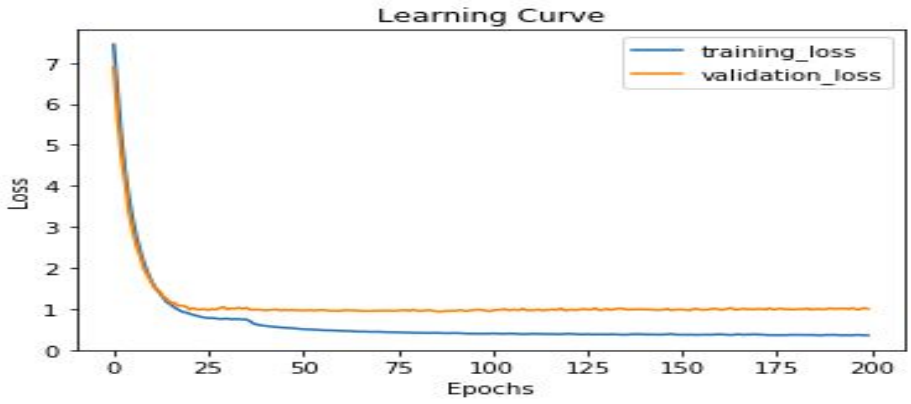
stream and mapped them to perform certain actions. The accuracy and loss plot for the same are given below :

```
In [0]: training_acc = hist.history['acc']
        val_acc = hist.history['val_acc']

        plt.plot(training_acc, label="training_accuracy")
        plt.plot(val_acc, label="validation_accuracy")
        plt.xlabel("Epochs")
        plt.ylabel("accuracy")
        plt.title("Learning Curve")
        plt.legend(loc='best')
        plt.show()
```



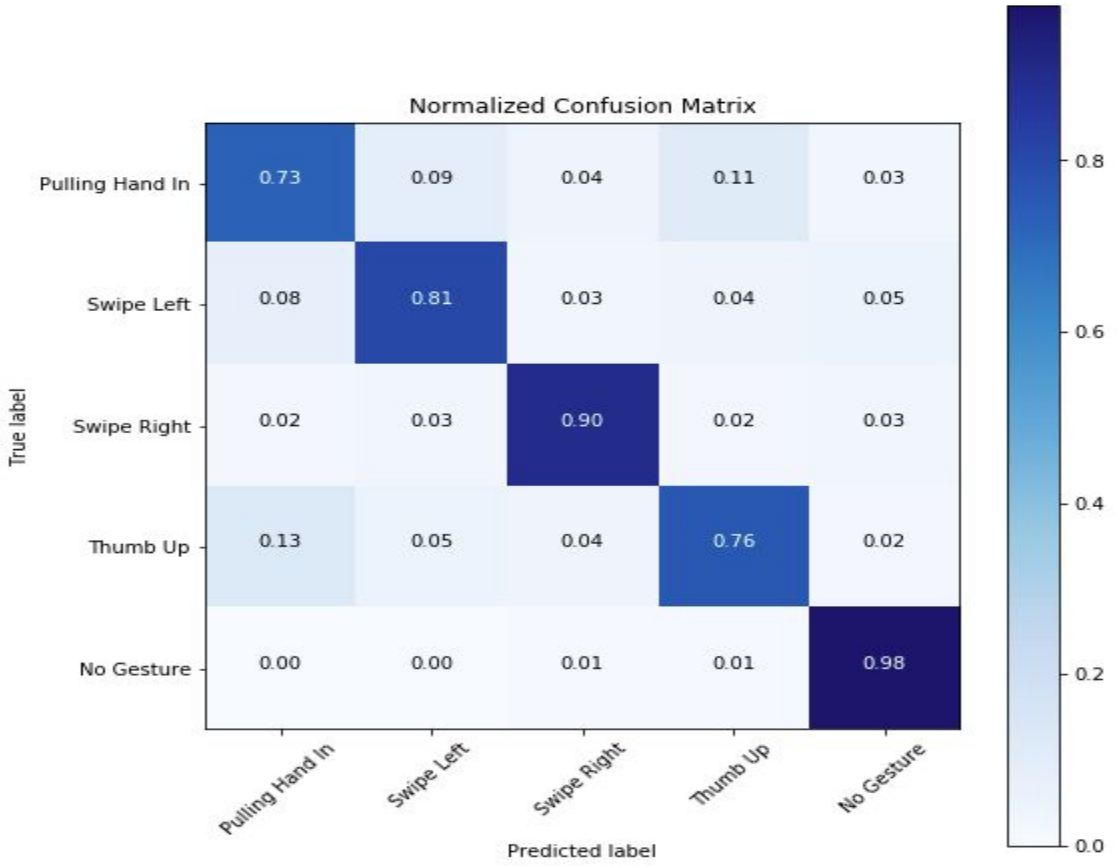
Accuracy Plot (Dynamic)

```
In [0]: training_loss = hist.history['loss']
        val_loss = hist.history['val_loss']

        plt.plot(training_loss, label="training_loss")
        plt.plot(val_loss, label="validation_loss")
        plt.xlabel("Epochs")
        plt.ylabel("Loss")
        plt.title("Learning Curve")
        plt.legend(loc='best')
        plt.show()
```



Loss plot (Dynamic)

The confusion matrix :



GRAPHICAL USER INTERFACE (GUI) :

A GUI or graphical user interface is a visual way of interacting with a computer using items such as windows, icons and menus used by most modern operating systems. A graphical user interface for the project was created using the python library Tkinter which loads the video stream and uses the loaded model to detect the gestures and also perform activities such as increasing the volume of the device with it.

CONCLUSION :

With the help of this study project, we were able to strengthen our grip on previously learnt concepts of Deep Learning and computer vision along with learning about the latest advancements and algorithms in the field. We were able to create an end-to-end pipeline for gesture recognition fused with human computer interaction which can be a stepping stone for further work in the field of advanced humanoid robotics and computer vision. The project gave us exposure to building deep learning models for both static and dynamic gesture recognition ,focusing on improving performance on real life data and GUI development. We sincerely thank our instructor Prof. Sujith Thomas for his constant guidance and support throughout the project.

REFERENCES :

1. https://arxiv.org/pdf/1901.10323.pdf
2. https://arxiv.org/pdf/1811.11997.pdf
3. https://medium.com/twentybn/gesture-recognition-using-end-to-end-learning-from-a-large-video-database-2ecbfb4659ff
4. https://gogul.dev/software/hand-gesture-recognition-p1
5. https://docs.python.org/3/library/tk.html