RUTGERS

Computer Vision: Project Phase 3

Name: Harish Udhaya Kumar
Course Instructor: Prof. Ahmed Elgammal

The scope of the project: Enhance existing sign language recognition's efficiency with a model as simple as possible, using as less parameters as possible, as less data as possible along with a high accuracy of sign language prediction.

Experiment:

In this phase, I have implemented three types of LSTM networks to test the accuracy and speed. The three networks are: Simple LSTM, Bidirectional LSTM and Stacked LSTM. Although the bidirectional LSTM and Stacked LSTM have their own advantages for this use case, a simple LSTM network gives a 99.99% accuracy in predicting the sign language. I experimented with other model architectures before selecting the presented model. I found that adding more than one LSTM or fully connected layer did not cause any notable difference in performance; thus, I removed these layers to minimize the model's capacity for overfitting. I also experimented with the output dimensionality of the LSTM: I tried 8, 16, 32, and 64. I found that using 32 and 64 performed similarly, with 64 usually performing slightly better.

Observations:

Below is the table of experimented models in LSTM network.

Model Type	Model Summary	Epochs	Accuracy
Simple LSTM	Total parameters: 562,115 LSTM Units: 64	60	99.9999%
Bidirectional LSTM	Total parameters: 1124227 Bidirectional units: 128	100	60%
Stacked LSTM	Total parameters: 716483 Stacked LSTM Units: (30 x 64), (30 x 120), 64, 64	100	98.99%

The tabulated results for 'Stacked LSTM' are close to the results generated by: Link

Datasets:
The data set that I used is not from an online source. Dataset used in this project is a live hand movement that I recorded from different people as part of the experiment.

The data recorded was a 30 frames/sequences for each sign and 30 such recordings. Therefore, the dimension of data for one sign would be (30, 30,L).
L = Length of the array for a sign (In our project it is 2132)

Enhancements (Project Phase 4 expectations):

The experiments that I would like to perform with this network is the following:

- Training the network in an ambience by varying the illuminance
- Currently, the "mediapipe" doesn't provide facial recognition feature. However, since there are approximately 458 landmarks on face, based on the dataset collected it is possible to evaluate the

coordinates of face and identify different people involved in the ambience. If the facial recognition is achieved, then the sign language recognition model would be least complex compared to existing models for generic activity recognition as well.