Dynamic Hand Gesture Recognition with Low Resolution Images

Ankit Kumar

Supervisor: Dr. Sunil Kumar

ABV-IIITM, Gwalior

September 25, 2022

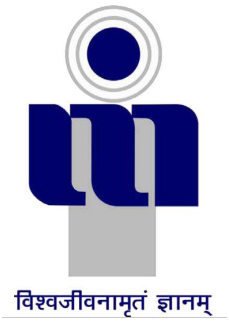

Table of Contents

1. Introduction
2. Literature Review
3. Objective
4. Methodology
5. Experimental Results
6. Discussion and conclusion
7. References

Introduction

- Recognition of hand gestures is crucial to the connection between the digital and physical worlds.
- The development of glove based approach was start of hand gesture recognition for computer control.
- There are two types of gestures as shown in Figure 1, that are classified on the basis of the movement.
- The difference between them is one is recognised using images and one uses videos.

Introduction

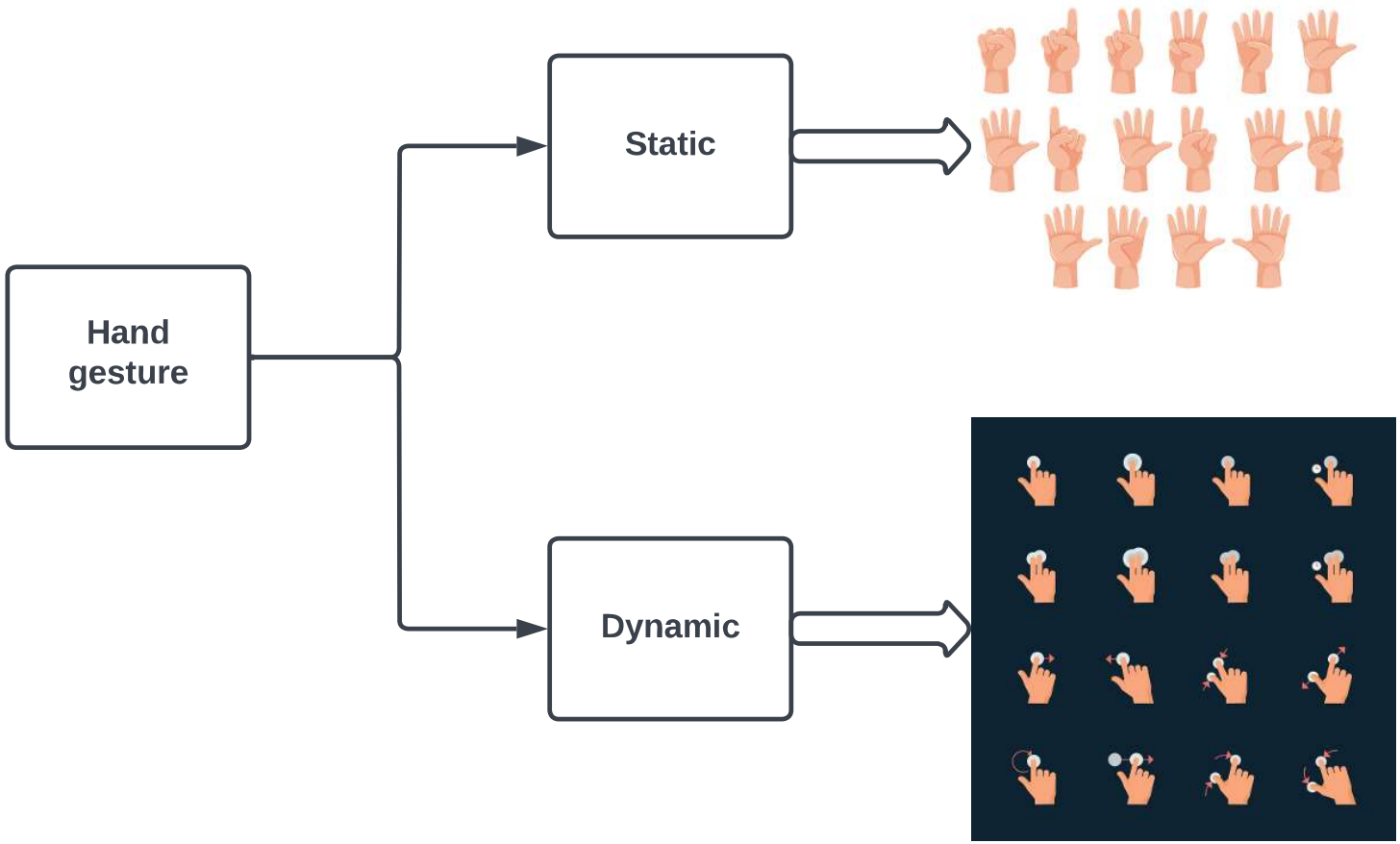

Figure 1: Classification of hand gestures

Introduction

- There are two types of approaches on which classification for hand gestures is based on.
- The first approach is based on vision, which uses a camera to acquire images and videos of hand movements over time.
- Another approach involves the usage of gloves which record the finger joint movements. It uses expensive data gloves.

Motivation

- Hand gestures are an inherent aspect of our interactions with the environment and are a crucial component of nonverbal communication.
- To break down the communication barrier with those who don't know sign language, developing hand motion detection technologies is crucial.
- When subject is a far from camera, recognition of hand gestures becomes difficult.
- Constructing hand gesture detection systems for distant subjects and weaker cameras inside of smart watches, which provide low-resolution pictures.

Research work flow

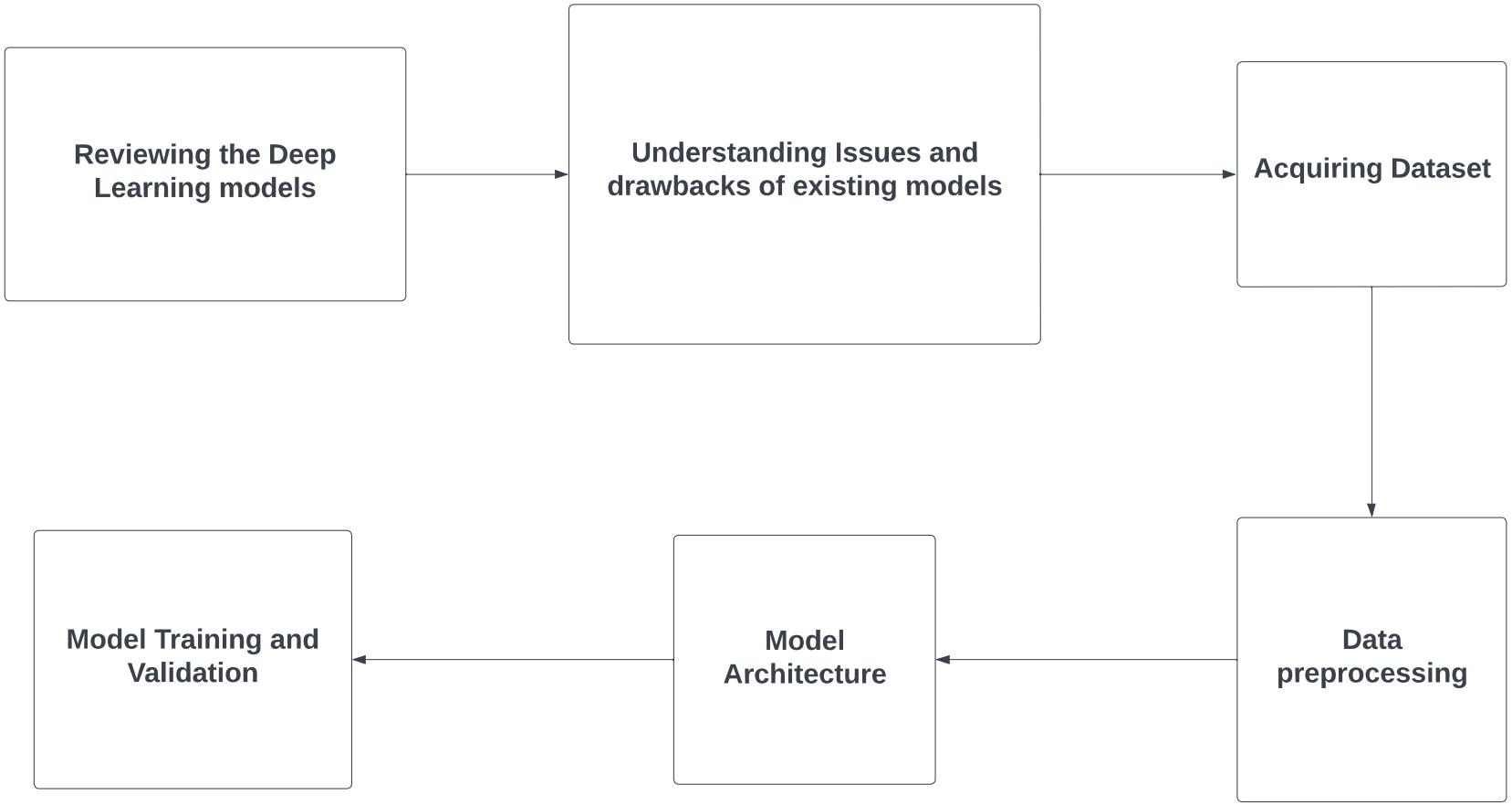

Figure 2: Activity flowchart

Literature Review

- Learning Spatio-temporal features is critical for performance to be stable in human hand gestures. Many methods have been proposed in recent years.
- As discussed by Munasinghe et al. proposed a feed-forward neural network-based approach for identifying four gestures [1].
- In this approach, as shown in figure 3, each frame is preprocessed. Each frame concatenate to form a single Motion History Image(MHI).
- When the deviation reaches a threshold value, then MHI is sent to the neural network for classification, which then returns probabilities for each type of gesture category.
- If the maximum probability exceeds 0.8, it is considered as a correctly classified gesture.

Literature Review

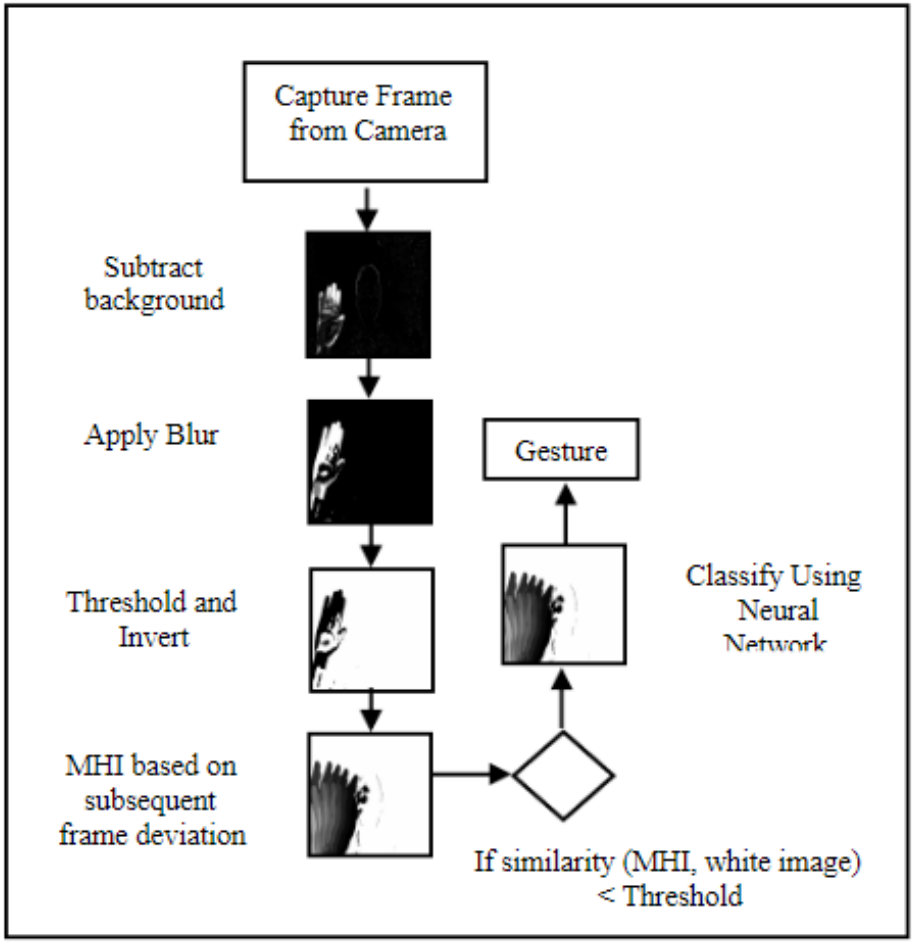

Figure 3: Overview of feed-forward neural network

Literature Review

- As discussed by Bao et al.[2], they proposed a two-dimensional nine-layer CNN model.
- It directly categorize hand gesture present in the images without pre-processing segmentation of the region of interest.

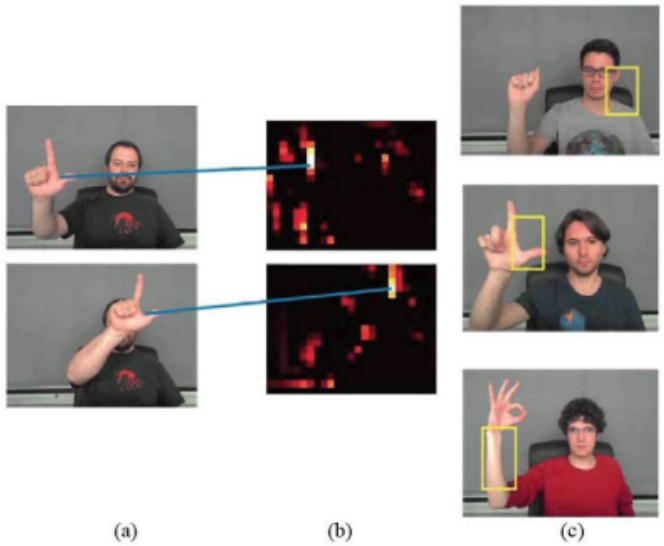

Figure 4: Detection of the same feature at different locations. Fig. (a) shows two images containing L-like shape hand gestures at different locations. Fig. (b) shows the corresponding feature maps. Fig. (c) shows examples where the model is detecting the L-like shape

Literature Review

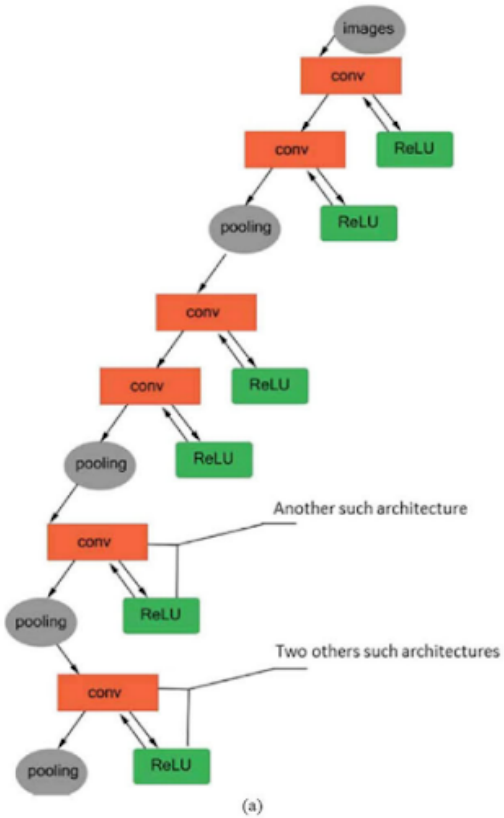

Figure 5: Architecture of 9 Layer CNN

Figure 6: Continuation of Architecture of 9 Layer CNN

Literature Review

- As discussed in [3], the architecture consists of a ResNet-18 which is 18 layers deep.
- ResNet-18 is used for extracting the features, followed by a transformer, which learns both spatial and temporal features of all video frames.
- Then using the softmax layer, gestures are classified.
- The use of depth and infrared images helped the model to detect gestures in low light.
- They also integrated multimodal in which two or more unimodal networks can be used simultaneously through the late fusion approach.

Literature Review

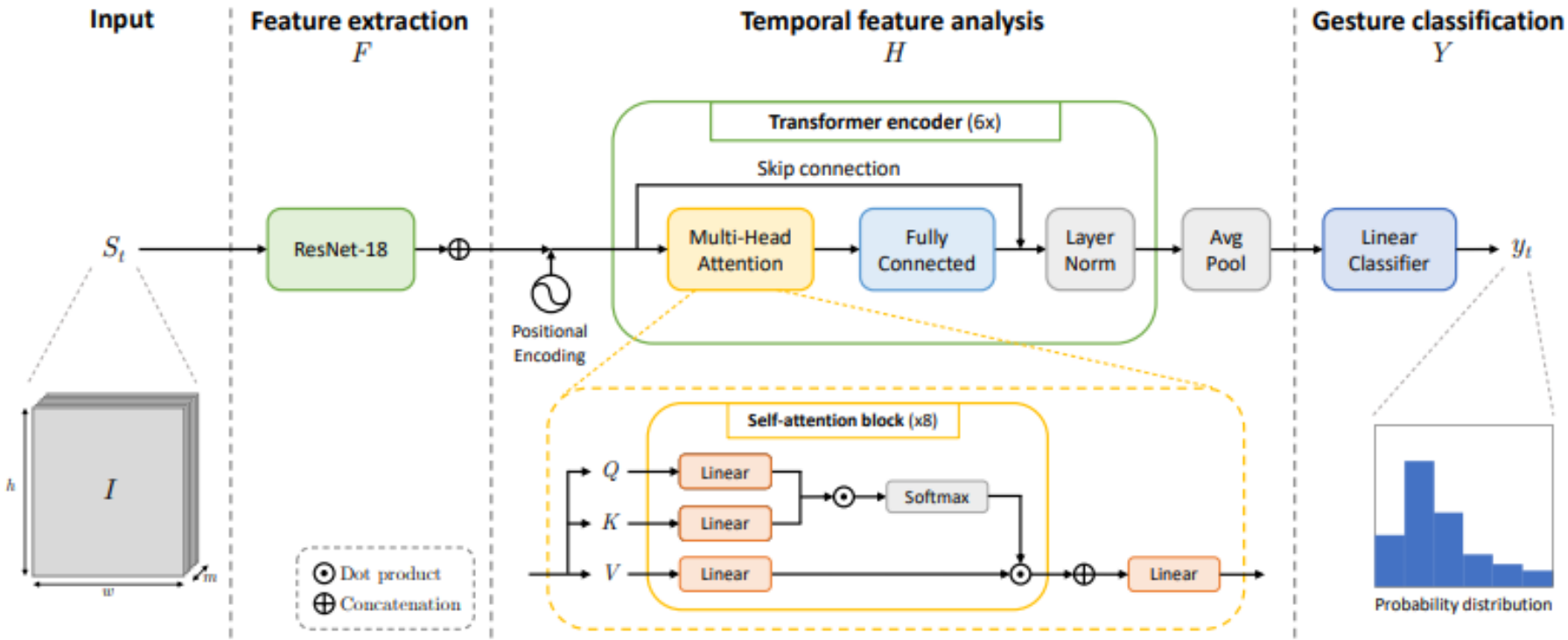

Figure 7: Overview of 3D-CNN+Transformer Model Pipeline

Literature Review

- As discussed in [4], the author proposed an R3DCNN for dynamic hand gesture recognition.
- The architecture consists of a deep 3D-CNN for spatio-temporal feature extraction, a recurrent layer for global temporal modelling.
- Then using the softmax layer, gestures are classified.

Literature Review

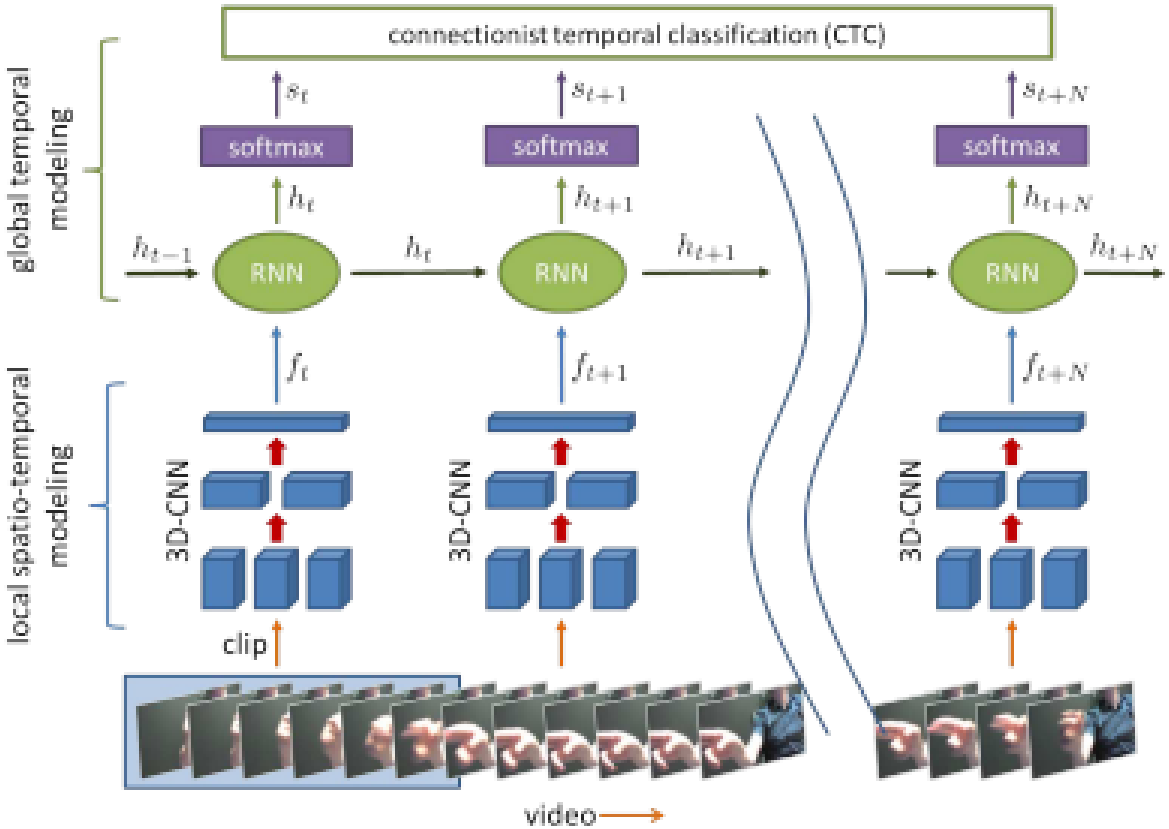

Figure 8: Overview of R3DCNN Model Pipeline

Research Gap

- Most works get confused between a hand and other parts of the arm.
- The detection of the beginning and ending gestures is not taken into account.
- Many methods do not distinguish between gesture and no gesture sequences.
- Very few research are carried out for low resolution videos.

Objective

- To come up with an architecture for dynamic hand gesture recognition for low resolution images.
- Address the temporal and sequential nature of dynamic gestures.

Proposed model

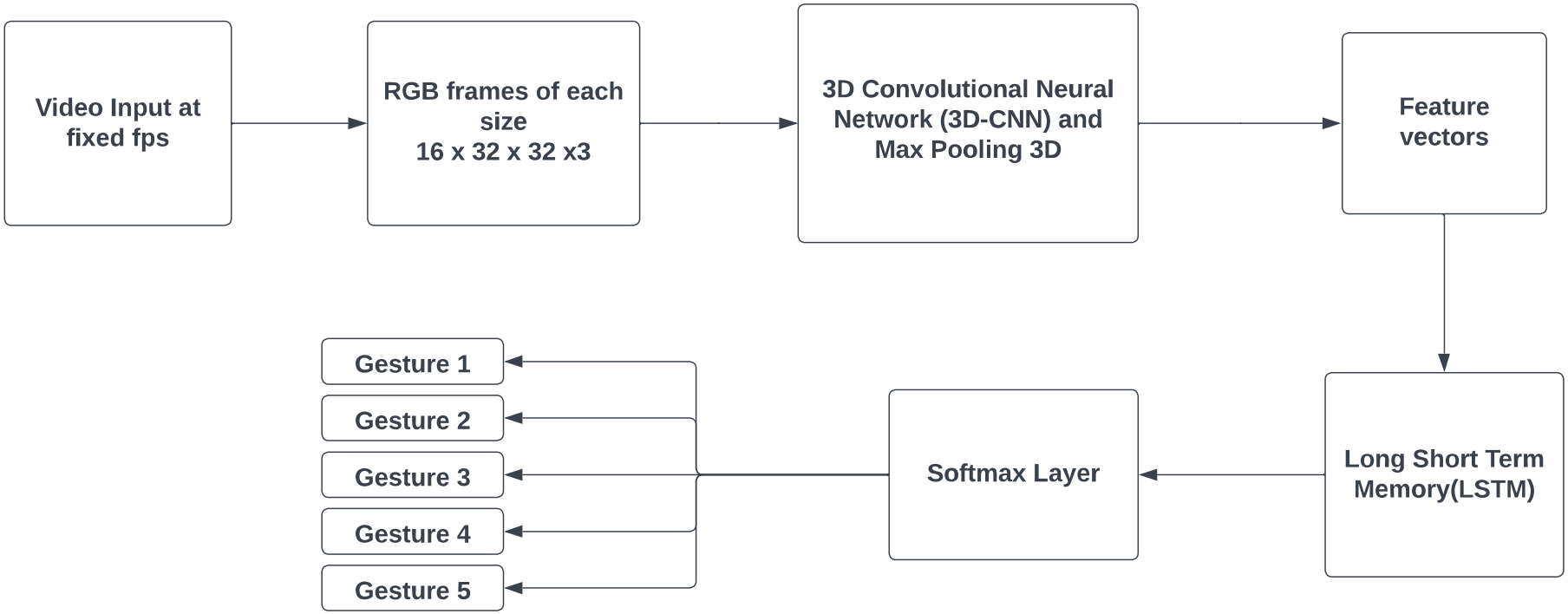

Figure 9: Overview of proposed model pipeline

Proposed model

- To address spatio-temporal information for gesture recognition, we proposed an architecture as shown in Figure 9 which consists of a combination of two networks 3D-CNN and LSTM, which is then given to softmax function for the classification.
- The work of 3DCNN algorithm employed is to extract the spatial data from subsequent video frames.
- 3DCNN generates a output which are feature maps that are converted into vector.
- LSTM network receive the input from 3DCNN with no. of samples and information of features which then do the categorization of hand motions, so that temporal information from sequence of images or video frames can be learned for long.

Data preprocessing

- We make a folder for each gesture and segregate them from other gestures,as all gestures are combined in a single folder.
- Two thousand movies are randomly selected for each class, and they are then split into a 80% training set and a 20% validation set.
- When importing data for training, every video frame is shrunk to a 32 by 32 pixel size.

Data preprocessing

Table 1: Splitting of dataset videos for training and validation

Total	**Training**	**Validation**
10,228	8,182	2,046

Feature Extraction

- The use of 3D-CNN network is to extract temporal data or features while maintaining the spatial data of the video frames(images).
- Usage of single network like 3D-CNN for dynamic gesture recognition to learn the long temporal and spatial data from video recordings is difficult.
- A new network that can learn for long temporal data is therefore required. An LSTM network and a 3D CNN are merged in our project and a basic general diagram is shown in Figure 10.

Feature Extraction

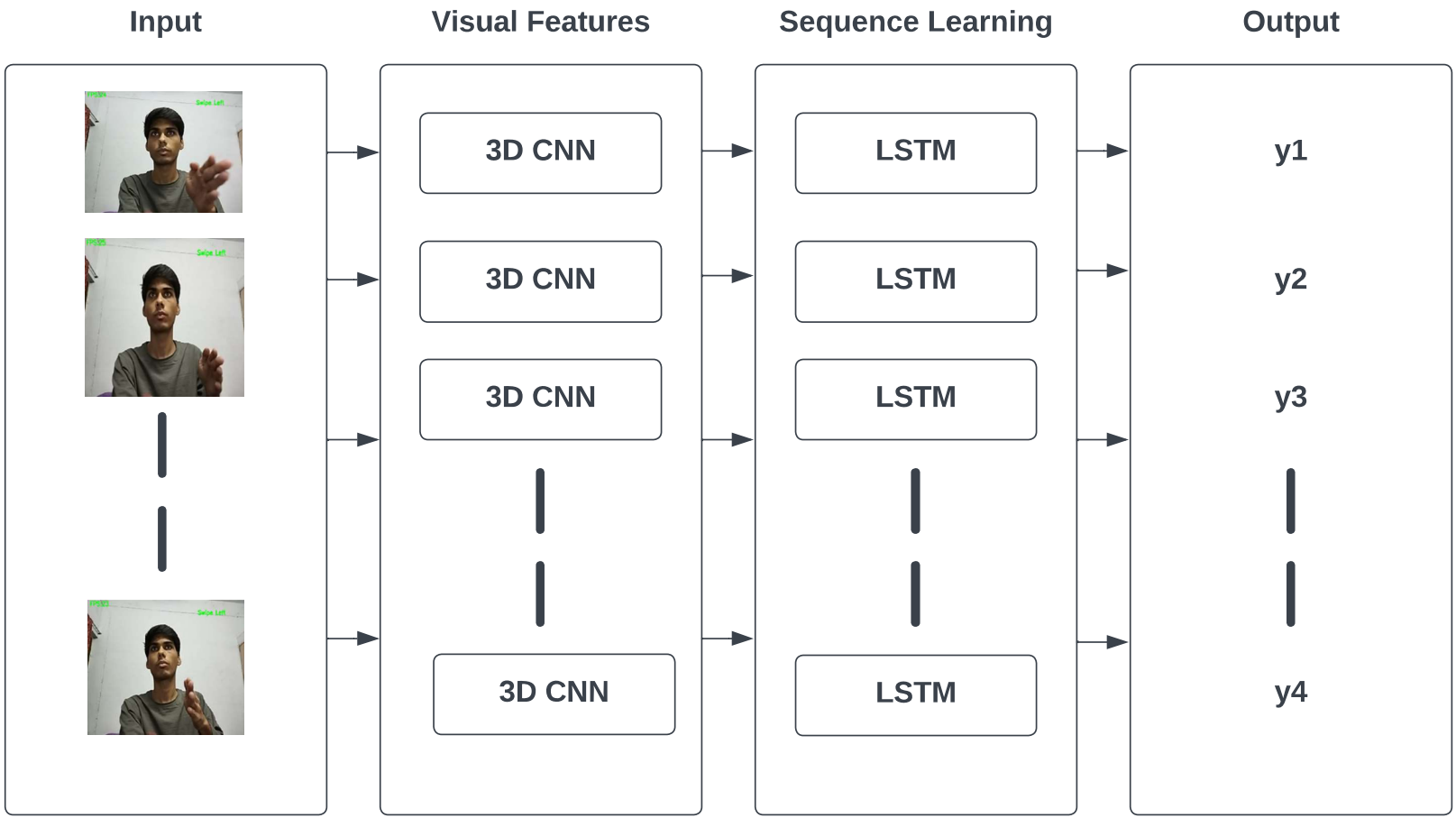

Figure 10: General diagram for proposed model

Feature Extraction

- We used three pooling layers and four convolution layers to learn each frame's visual attributes.
- The batchnorm layer transmits the characteristics collected from the three dimensional convolution neural network to the LSTM.
- With the exception of the first layer, which has a size of 1x2x2, every Conv3D layer has a pooling size of 2x2x2 and a kernel size of 3x3x3.
- This layer preserves the temporal properties.
- We used dense layer to connect the model and dropout layer to reduce overfitting which randomly drop out neurons.

Feature Extraction

Table 2: Layer-wise details of the proposed architecture for hand gesture recognition

No.	Layer	Filters/Pooling
1	Conv 3D 3x3x3, ReLu	8
2	Max Pooling 3D	(1,2,2)
3	Conv 3D 3x3x3, ReLu	16
4	Max Pooling 3D	(2,2,2)
5	Conv 3D 3x3x3, ReLu	32
6	Conv 3D 3x3x3, ReLu	32
7	Max Pooling 3D	(2,2,2)
8	Batch Normalizaton	–
9	LSTM	32
10	Flatten	–
11	Dense, ReLu	512
12	Dropout	–
13	Dense, ReLu	256
14	Dropout	–
15	Softmax	–

Experimental Setup

1. Dataset used in our project is 20BN-Jester.
2. We use only 5 gestures out of 27 hand gestures due to memory resource constraints.
3. We choose 2000 videos at random per class out of more than 3000 videos, which is further divded into 80% training and 20% validation.
4. Adam optimizer is used.
5. Categorical crossentropy loss function is used.
6. Batch Size = 32
7. Frame size = 16, Image size = 32x32
8. Epoch = 70

Experiment 1

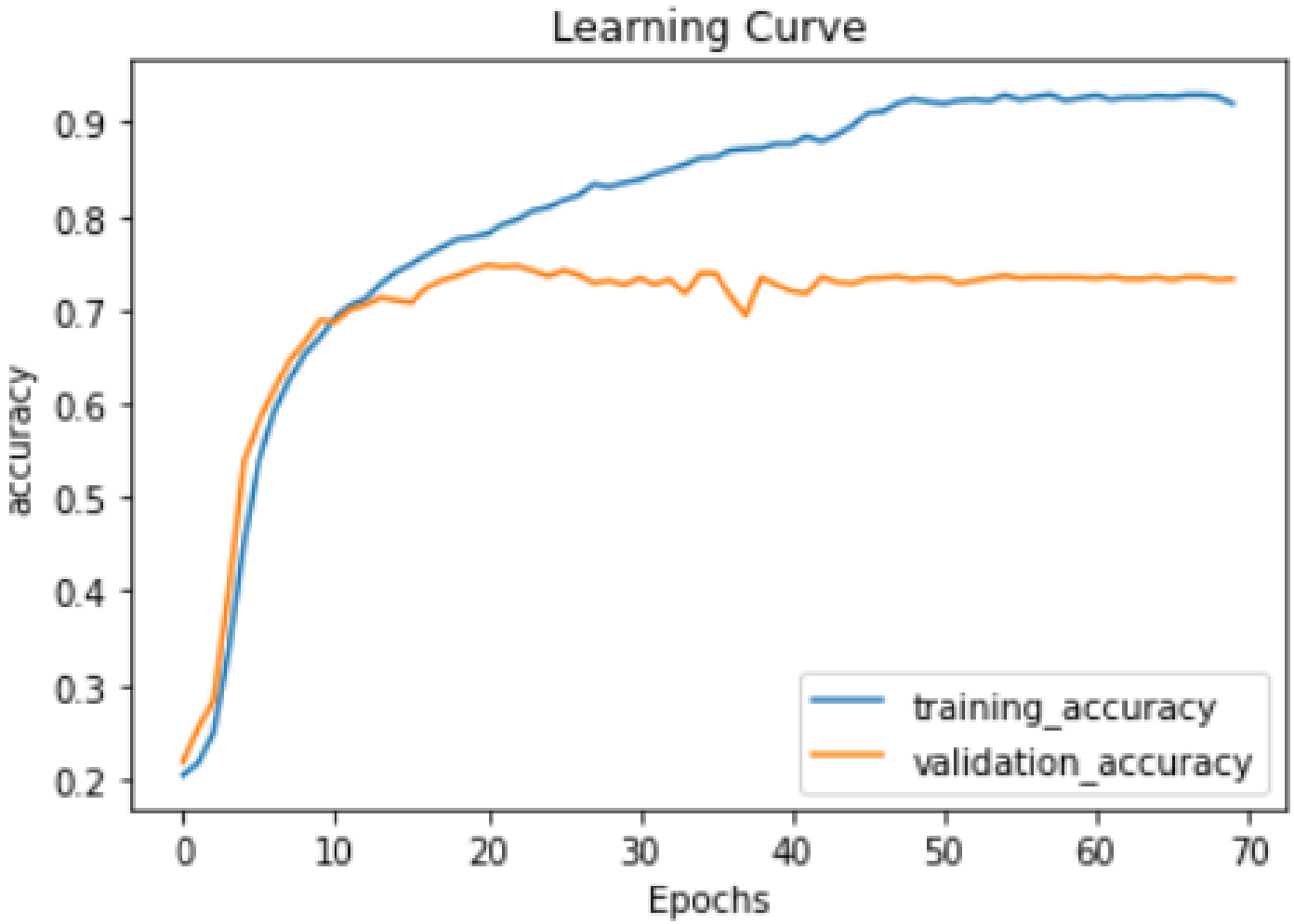

Figure 11: MobileNet-V2 + LSTM model accuracy

Experiment 1

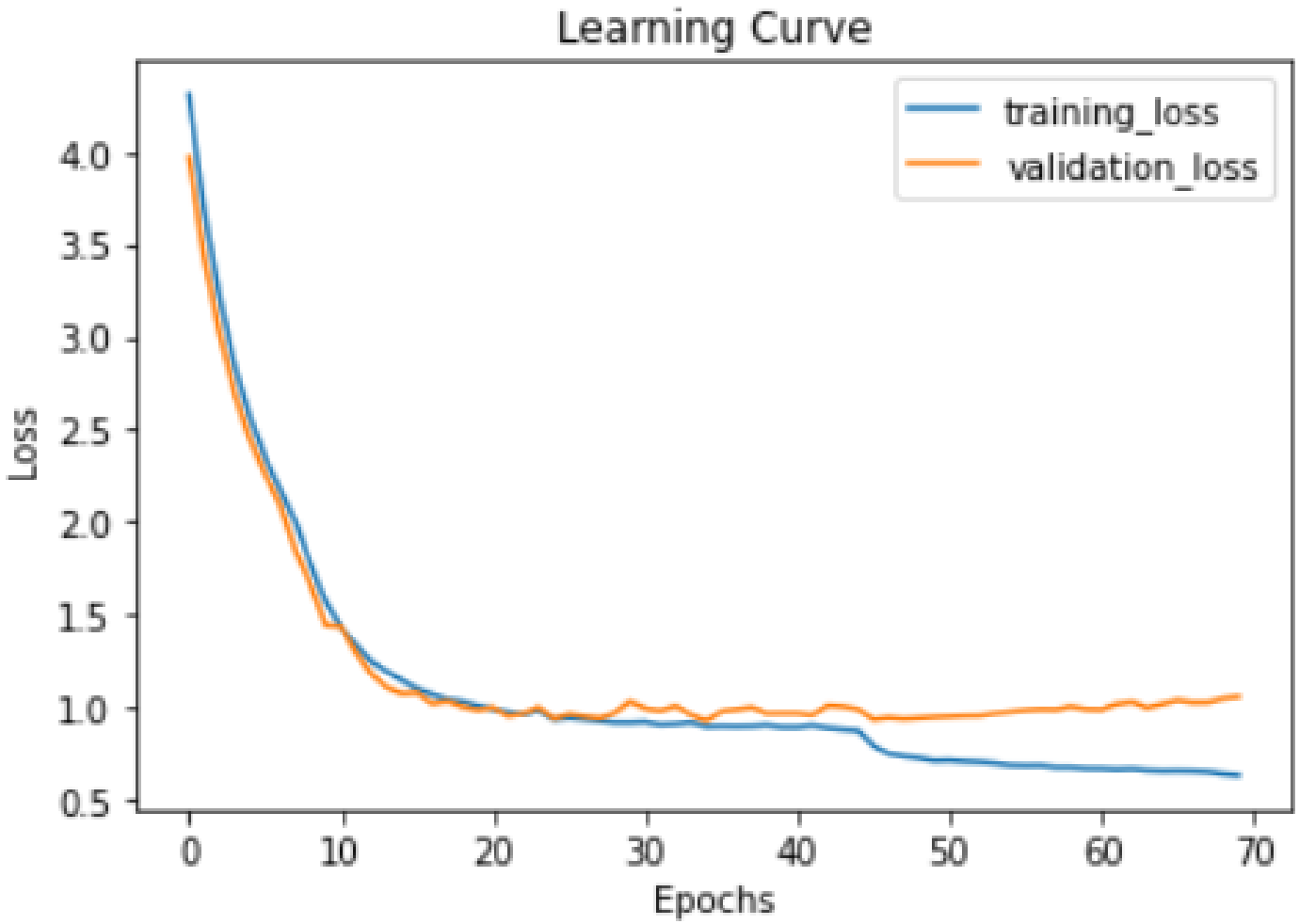

Figure 12: MobileNet-V2 + LSTM model loss

Experiment 2

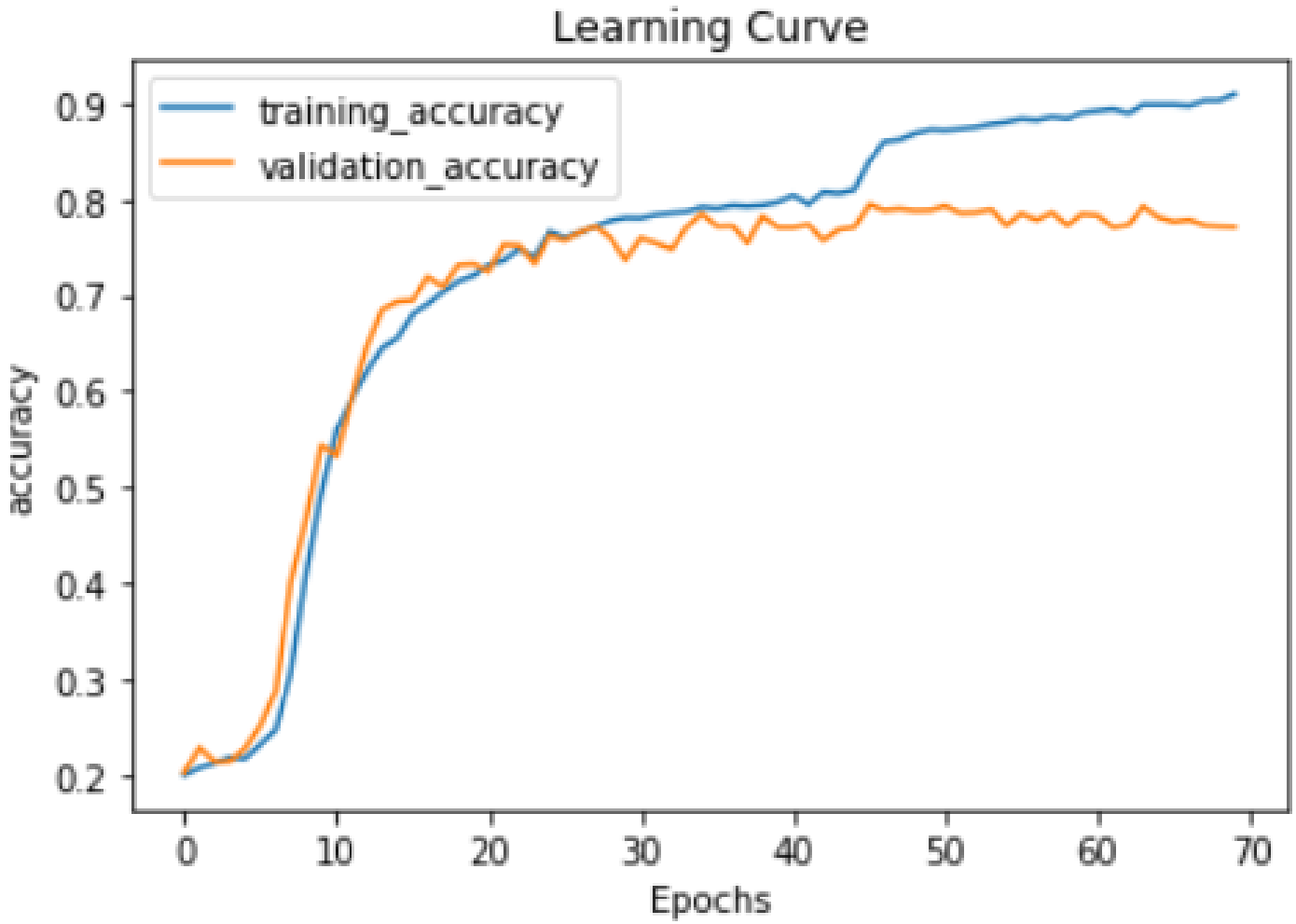

Figure 13: MobileNet-V2 + LSTM + batchnorm model accuracy

Experiment 2

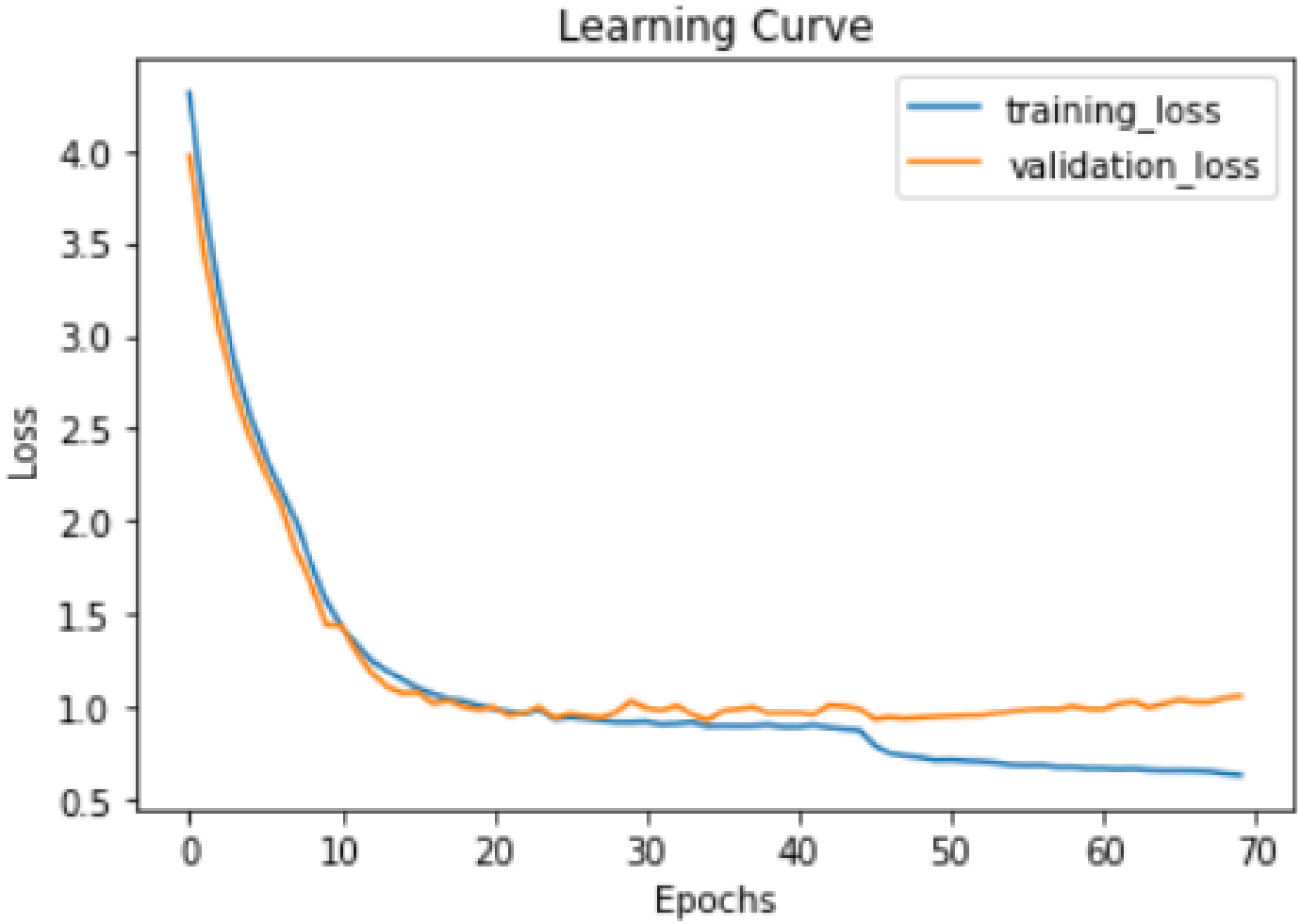

Figure 14: MobileNet-V2 + LSTM + batchnorm model loss

Experiment 3

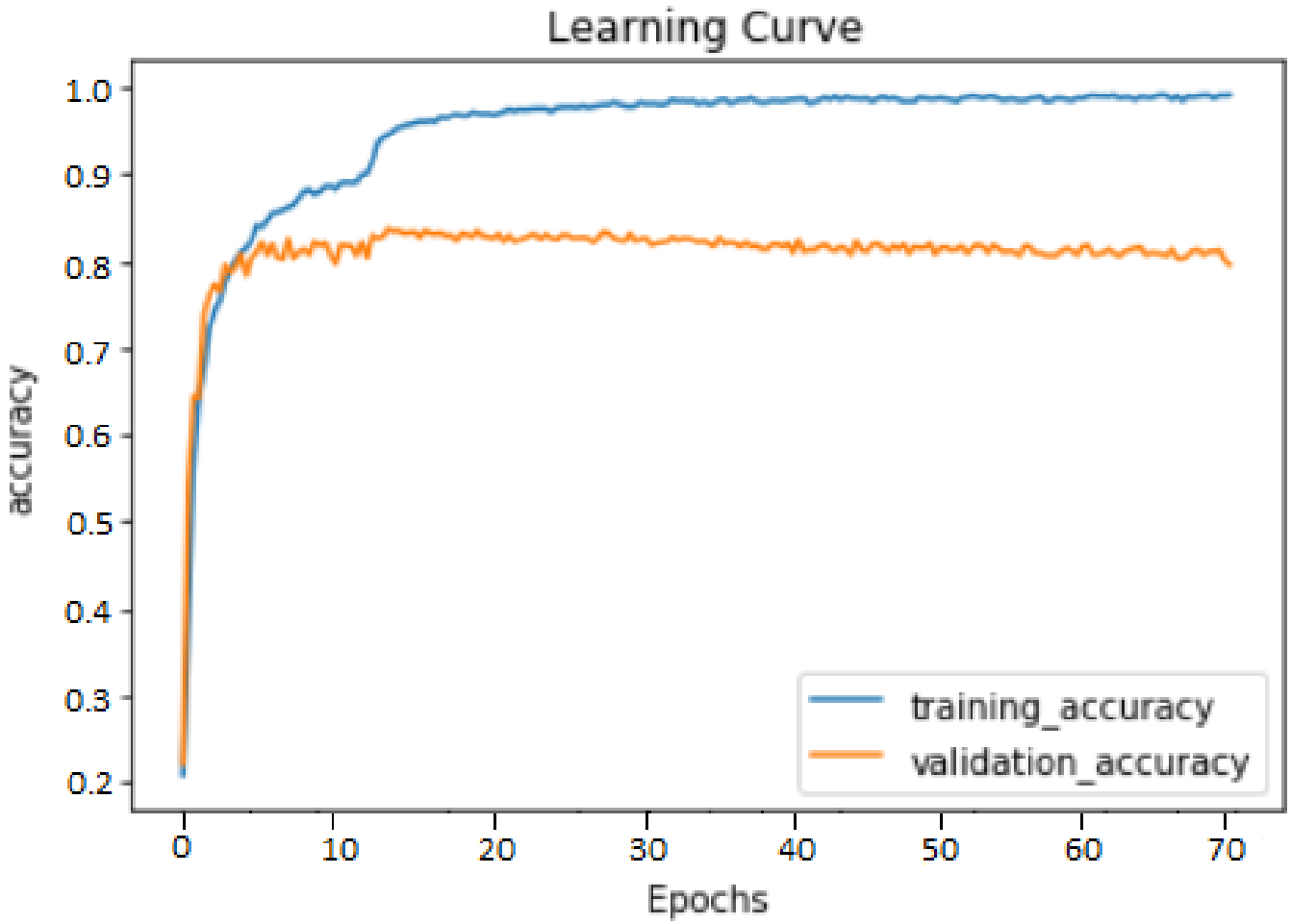

Figure 15: Accuracy of proposed model

Experiment 3

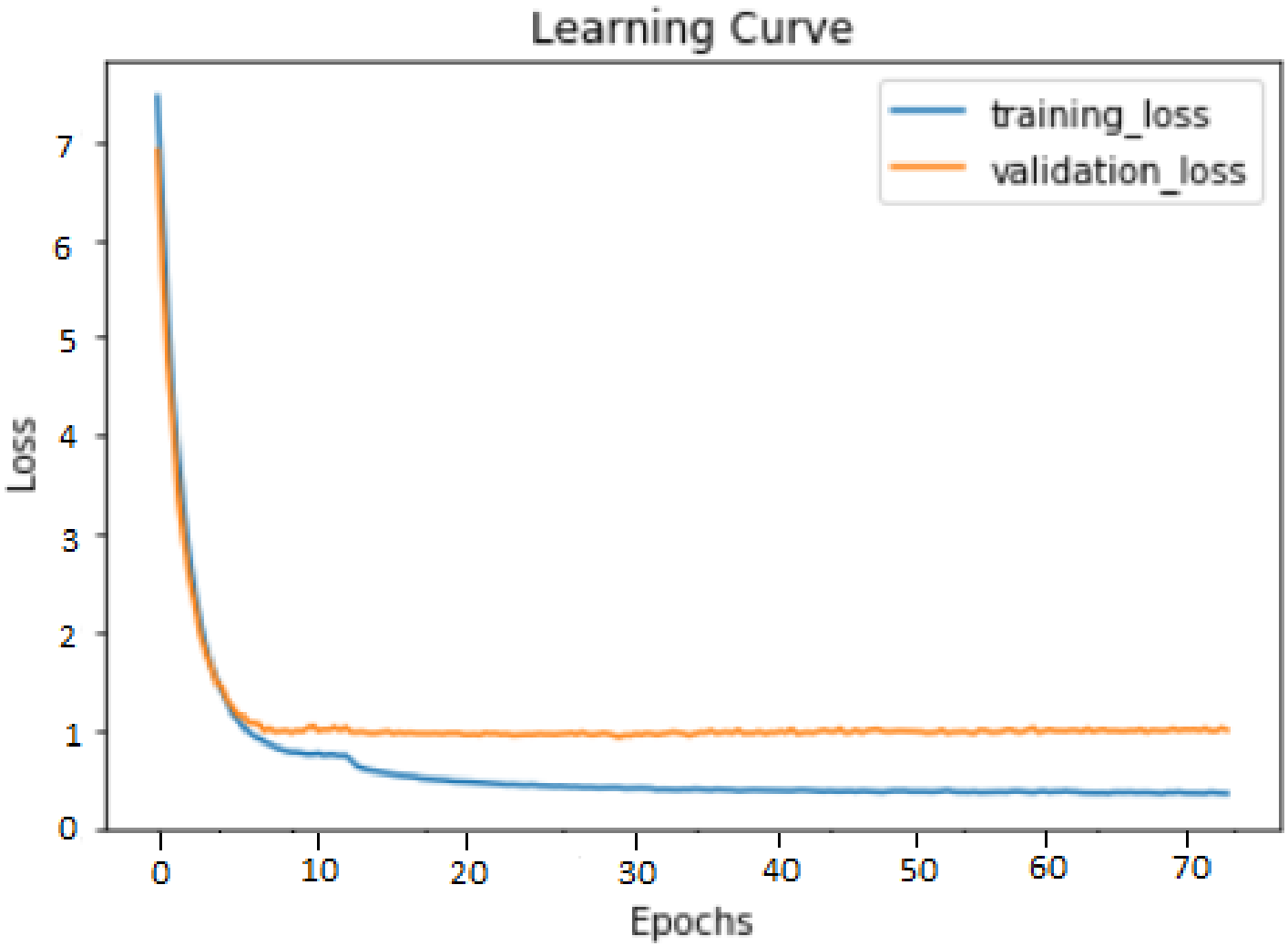

Figure 16: Proposed model loss

Experimental results

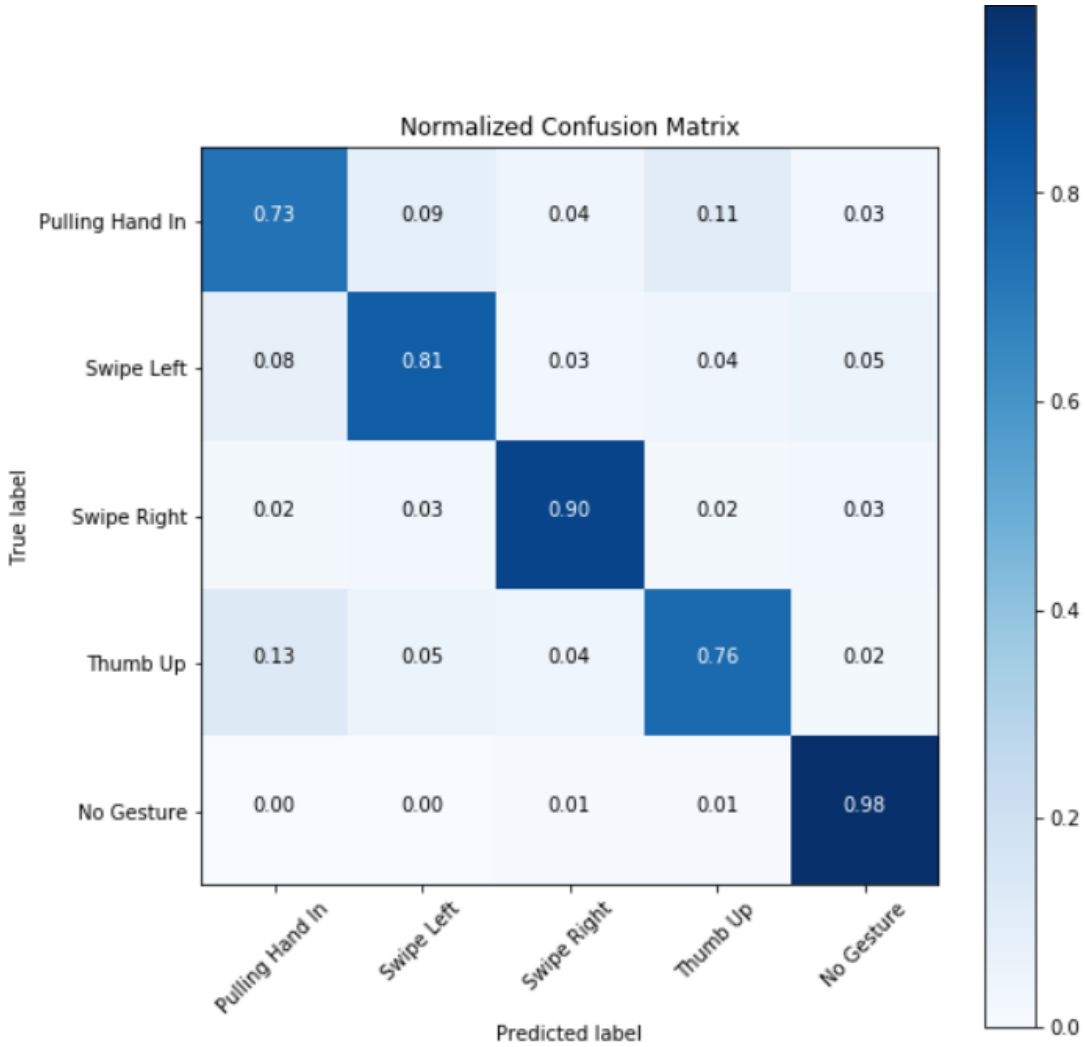

Figure 17: Confusion Matrix for proposed model on test dataset

Experimental Conclusion

Table 3: Comparison of different models discussed in experiments

Model	**Training Accuracy**	**Validation Accuracy**
MobileNet-V2 + LSTM [5]	91%	72%
MobileNet-V2 + LSTM + batchnorm [5]	92%	77%
3DCNN + LSTM (Proposed model)	98%	83%

Conclusion

- Learning of both temporal and spatial aspects is done by 3DCNN + LSTM for all sequence of images under challenging backdrop, illumination conditions and low resolution video.
- In comparison to MobileNetv2 +LSTM, the proposed model consisting of 3DCNN + LSTM produces better results.
- During testing, **82%** accuracy was achieved for our proposed model.

Challenges and future work

Challenges faced during the project.

- Dataset : Qualcomm purchased the dataset we utilised, which included images of the videos in different configurations.
- Hardware Limitations : There is limited amount of resources available for training of model.

Future work

- More effective and advanced deep learning techniques can be applied.
- The model can be expanded with more gesture classes.
- Building of an application for dynamic had gesture recognition, which can be utilised for HCI.

References I

[1] N. Munasinghe, "Dynamic hand gesture recognition using computer vision and neural networks," 04 2018.

[2] P. Bao, A. I. Maqueda, C. R. del Blanco, and N. García, "Tiny hand gesture recognition without localization via a deep convolutional network," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 3, pp. 251–257, 2017.

[3] A. D'Eusanio, A. Simoni, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara, "A transformer-based network for dynamic hand gesture recognition," in *International Conference on 3D Vision*, 2020.

[4] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, 09 2018.

[6] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2874–2882, 2019.

[7] T.-D. Truong, Q.-H. Bui, C. N. Duong, H.-S. Seo, S. L. Phung, X. Li, and K. Luu, "Direcformer: A directed attention in transformer approach to robust action recognition," in *Computer Vision and Pattern Recognition*, 2022.

Thank You