Hybrid CNN-SVM model for Action Recognition from Videos

Submitted By :-
Name:-Anik Mitra (10200119064)
Name:-Uttam Modi (10200119058)
Name:-Pratik Tamang (10200119066)
Name:-Rahul Pramanik (10200119061)

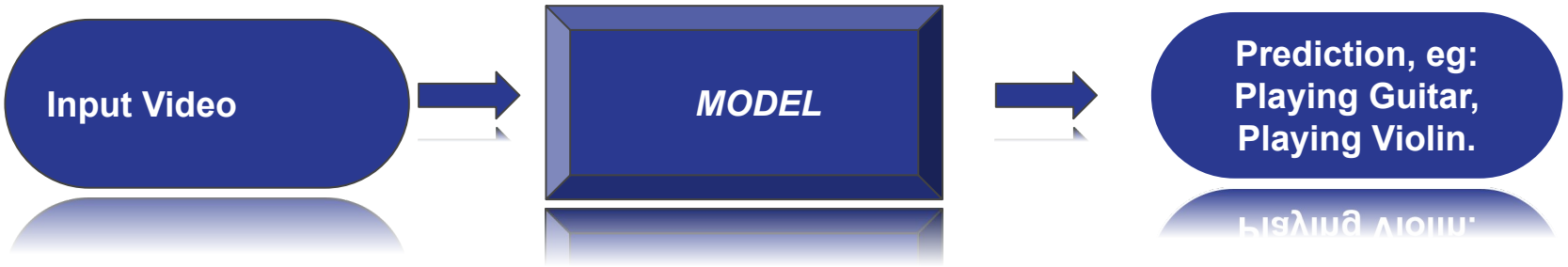
Under the guidance of
Dr.Kousik Dasgupta
Computer Science and Engineering
Kalyani Government Engineering College

CONTENTS

INTRODUCTION

Action recognition is the task of identifying an action in a video by Neural Network or Artificial Intelligence model.

- Goal: To train a Artificial Intelligent Model that can analyze a video and able to predict what action is taking place with efficiently and good accuracy.



LITERATURE SURVEY

[1] The author developed a deep convolutional network armature for detecting mortal conduct in videos by using the action bank features of the UCF50 database.

[2] the author established 3D CNN models for action recognition .These models develop features from both spatial and temporal measurements by performing 3D convolutions

[3] The author analyzes mortal action recognition as being a grueling task due to both the spatial and temporal confines of the action data.

PROPOSED MODEL

- CNN(Convolutional Neural Network) + LSTM (Long Short Term Memory)
- CNN(Convolutional Neural Network) + SVM(Support Vector Machine).

Libraries

Keras

Tensorflow

Dataset

UCF-101

DATASET CLASSES

- UCF101 dataset is an extension of UCF50.
- Categories of UCF101.
- Total video hours of Dataset .
- FPS & resolution of the videos.
- Playing Cello, Playing Guitar Playing Dhol, Playing Flute, Playing Piano, Playing Sitar, Playing Tabla' Playing Violin, Playing Daf, Drumming

PREPROCESS OF DATA

- We resize the frames of the videos to a fixed width and height.
- To reduce the computations and normalize the data to range.
- The range will be [0-1] by dividing the pixel values with 255, which makes convergence faster while training the network.
- We have taken the number of frames of a video that will be fed to the model as one sequence will be 20 so that we get the whole idea of what is happening in the video.

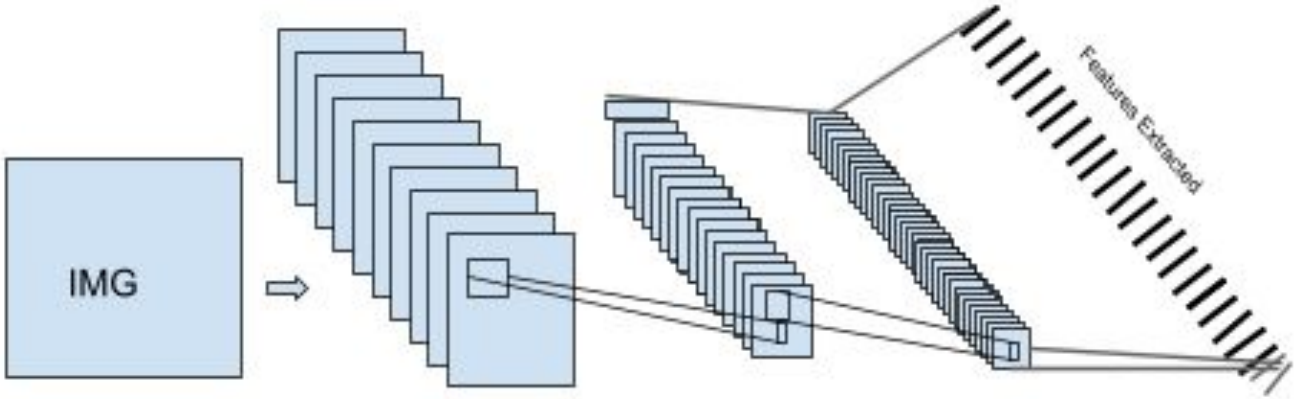
- Frame Extraction
- Data normalization

CONVOLUTIONAL NEURAL NETWORK

Architecture of Convolutional Neural Network

- **Convolutional Layer:**
- **Pooling Layer:**
- **Fully-Connected layer:**

CNN is a neural network that **extracts spatial features**



LONG SHORT TERM MEMORY (LSTM)

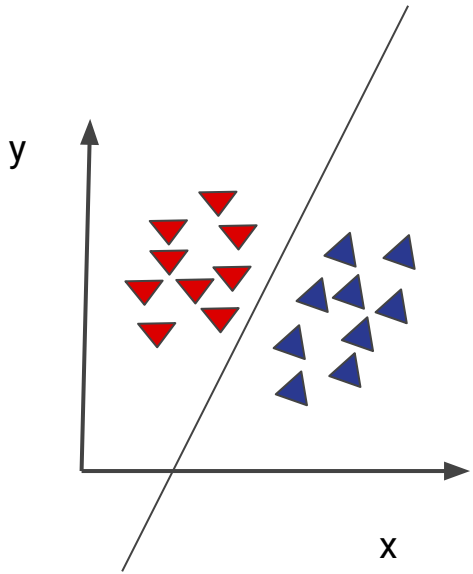
- It work very well in sequential data like audio, video and text.
- LSTM used for Classification problems.
- LSTM models with Convolutional Neural Networks excel at learning spatial relationships.

Proposed HYBRID CNN WITH LSTM model

- We have used time-distributed Conv2D layers which will be followed by MaxPooling2D.
- The feature extracted from the Conv2D layers will be then flattened using the Flatten layer and will be fed to a LSTM layer.
- The Dense layer with softmax activation will then use the output from the LSTM layer to predict the action being performed.

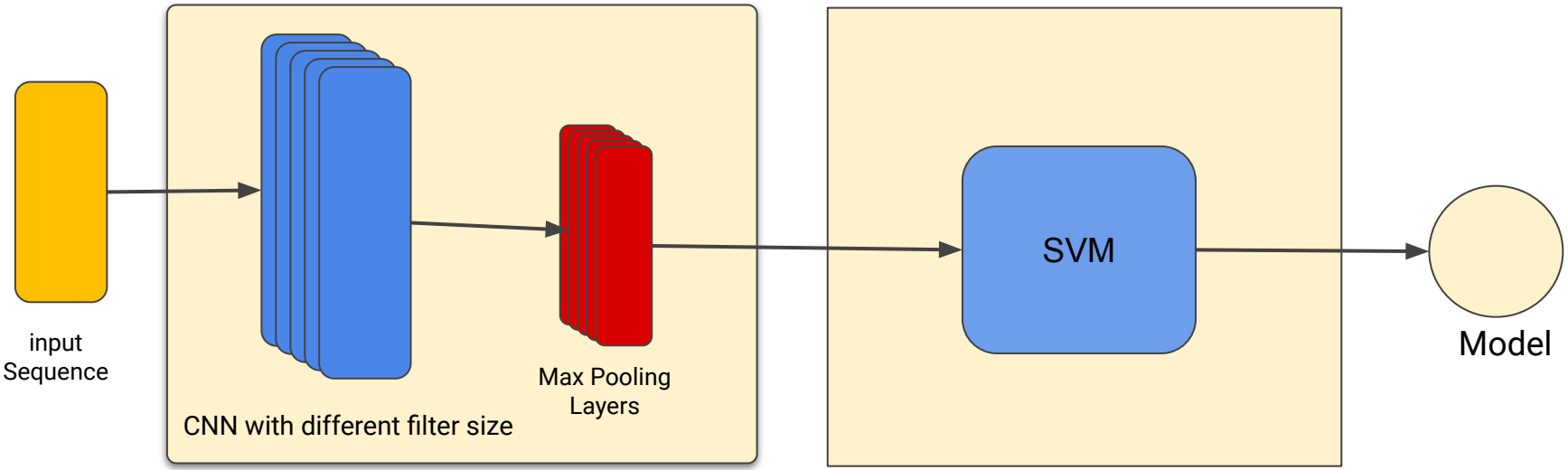
SVM(Support Vector Machine)

- Support Vector Machines(SVM) is considered to be a classification approach but it can be employed in both types of classification and regression problems.
- It's a supervised learning algorithm that is mainly used to classify data into different classes.
- SVM segregate the given dataset in the best possible way.
- In our model for cnn+svm features from CNN fed into SVM algorithm for video classifier we use loss function ' squared hinge' for multiple classes and activation function 'softmax'.

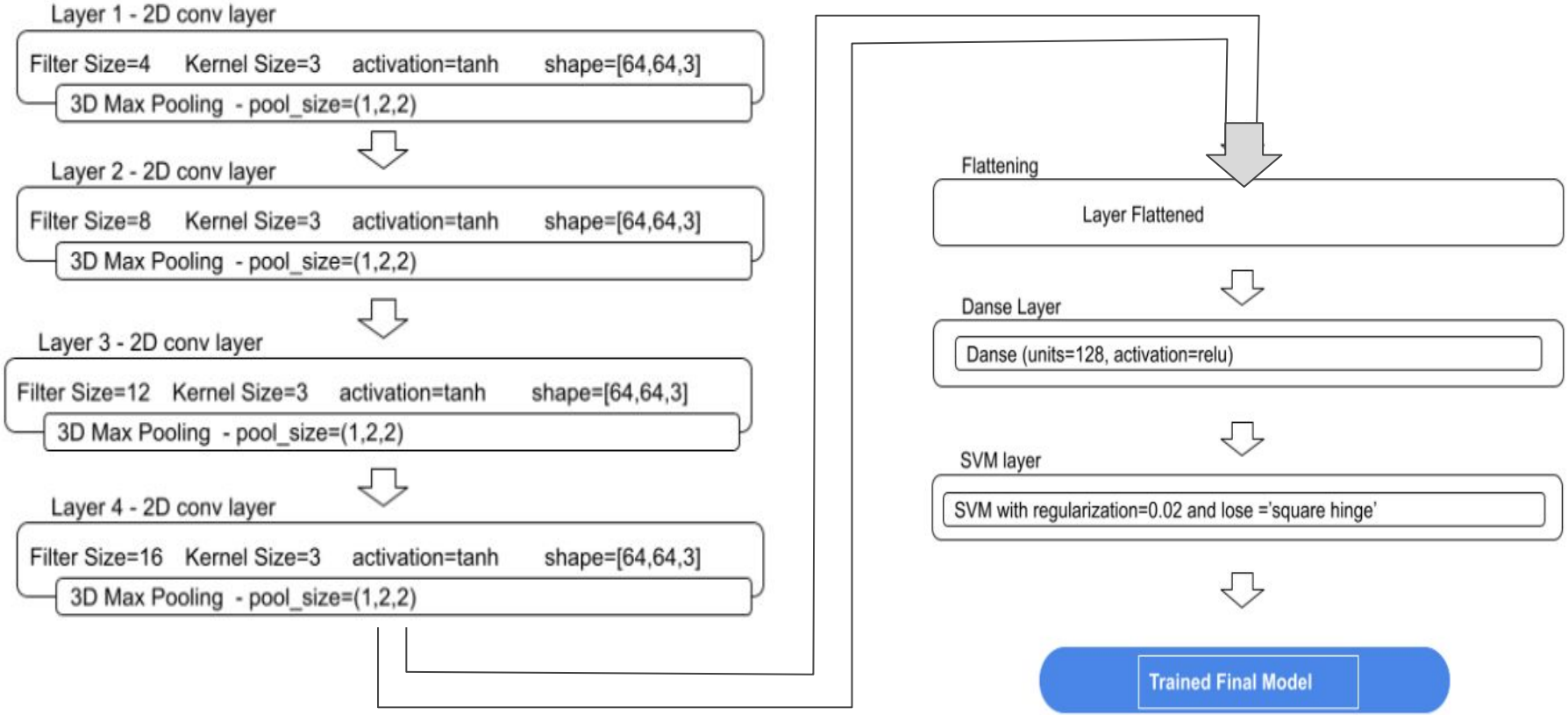


SVM

A NOVEL CNN + SUPPORT VECTOR MACHINE(SVM) APPROACH FOR ACTION RECOGNITION IN VIDEOS



The LAYERS OF FINAL CNN + SVM MODEL



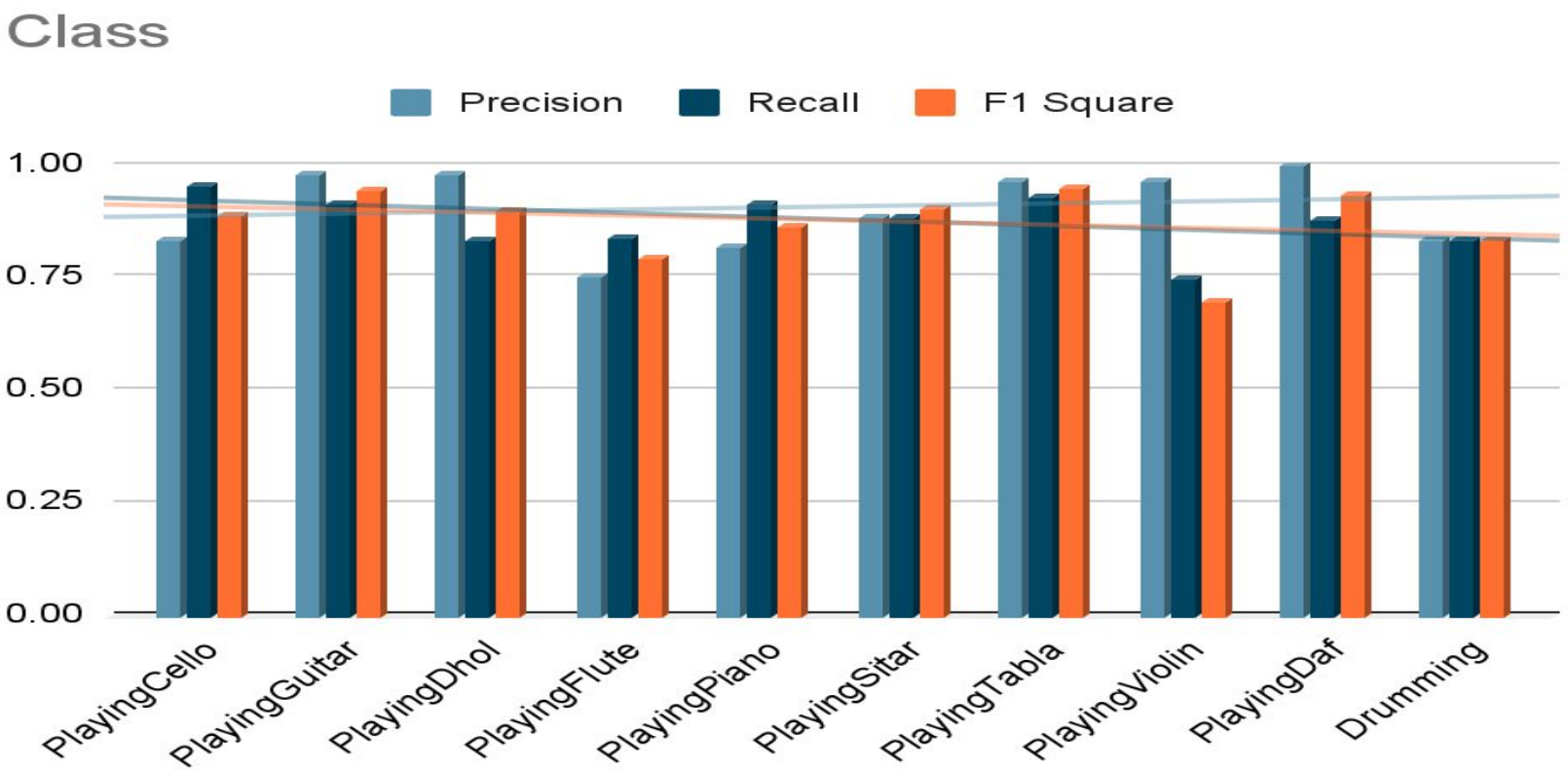
DIFFERENCE BETWEEN SVM AND LSTM AS CLASSIFIER

Structure: SVM possesses a number of parameters that increase linearly with the linear increase in the size of the input. LSTM, on the other hand, doesn't.

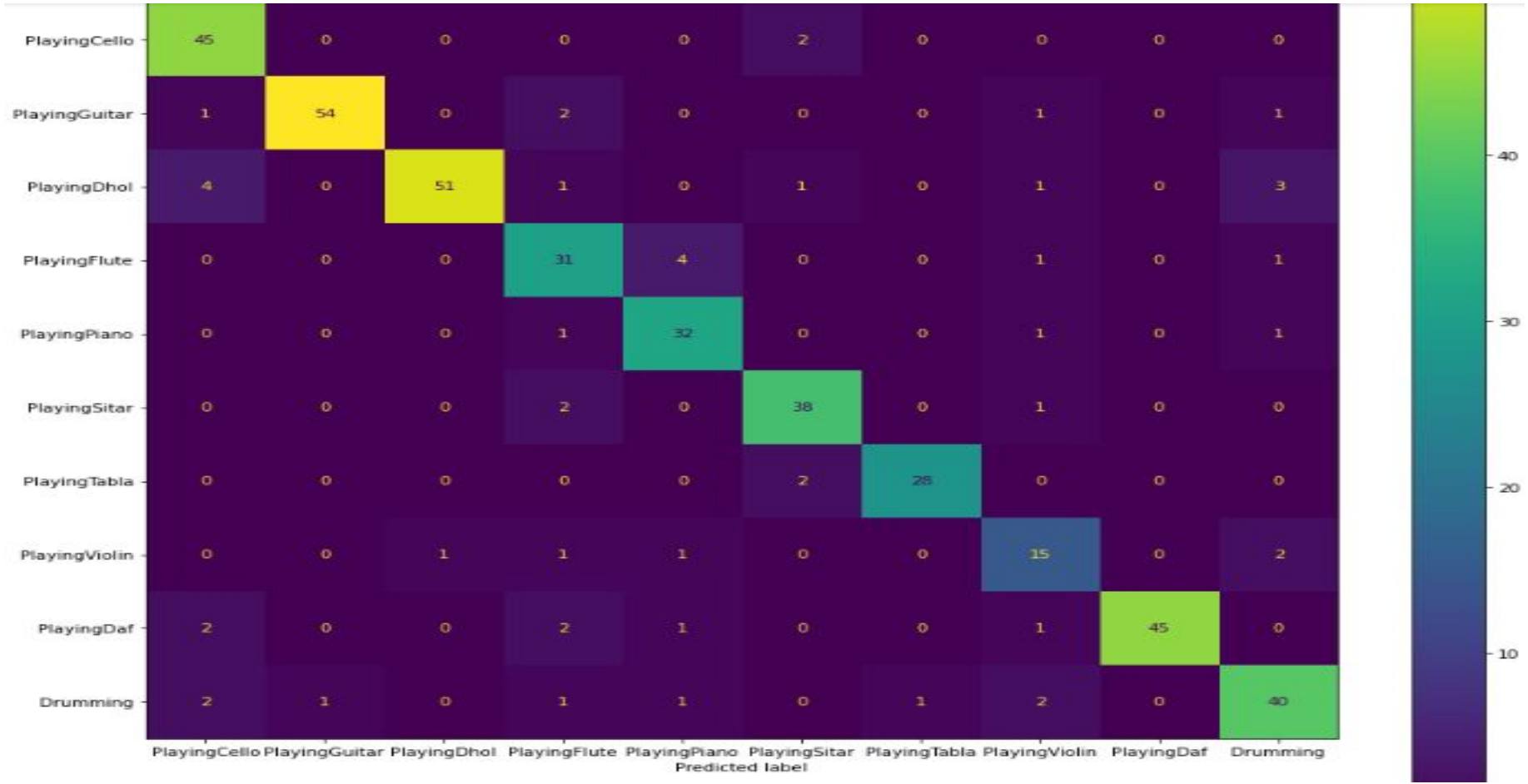
Required of Training Data: Support vector machines effectively use only a subset of a dataset as training data. Which reduces the training data, in turn, LSTM requires huge training dataset.

Training time for Algorithm: SVMs are generally very fast to train, while lstm is slow.

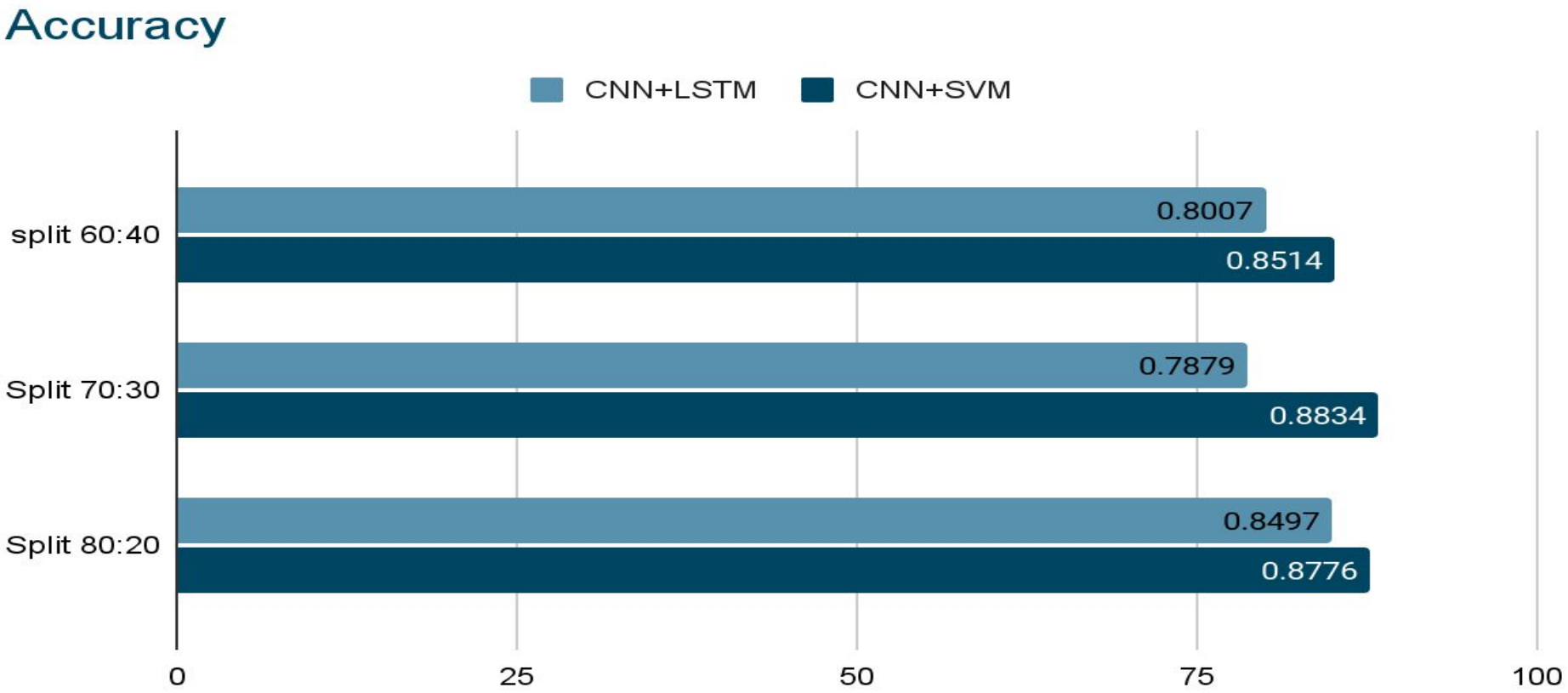
RESULT AND ANALYSIS



CONFUSION MATRIX OF OUR PROPOSED MODEL



PERFORMANCE EVALUATION OF THE PROPOSED METHOD WITH OTHER METHODS BASED ON SPLIT RATIO.

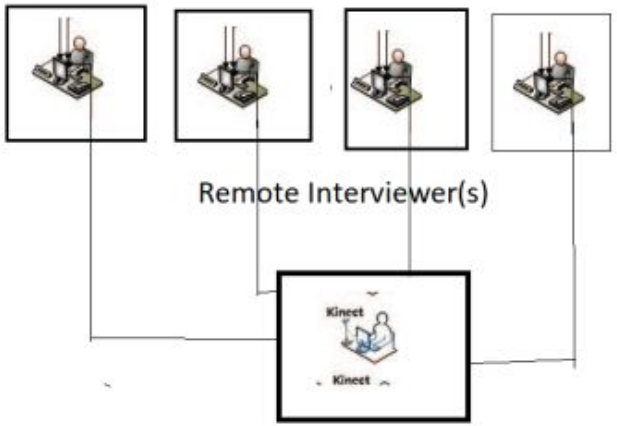


PREDICTED OUTPUT GENERATED



APPLICATION OF PROPOSED WORK

- In our Model we have taken some music related classes where we can identify what instrument the person is playing or trying to play.
- The applications range from surveillance in Industry line, online interview for job aspirants or may be Cheating on the Seat of The Examination.
- The presented work is a part of a surveillance system for remote online based Interview systems to facilitate recruiters.



Remote Candidate being monitored for illegal activity or confidence level of the candidate

CONCLUSION

- We carried out a brief study of Hybride CNN-SVM Model for Action Recognition from videos and proposed some experiments on it.
- We took the **UCF101** dataset and with the 10 classes/categories.
- Preprocessed the dataset and took out the features from the series of images using **CNN.**
- Splitted the dataset into various split ratio like **60:40 , 70:30** and **80:20.**
- For the Classification of actions we used LSTM and gained an accuracy of **78.9** % with **70:30(standard split ratio).**
- Furthermore, for better classification we moved to CNN with Support Vector Machine**(our Proposed model)** from LSTM with an accuracy of **88.34 %**

FUTURE SCOPE

Action Recognition is an important problem in computer vision. AR is the basis for many applications such as video surveillance, health care, and human-computer interaction. Methodologies and technologies have made tremendous development in the past decades and have kept developing up to date. However, challenges still exist when facing realistic sceneries.

REFERENCE

[1] S.Sadanand and J.Corso,“Action bank: A high-level representation of activity in Video,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition ,pp. 1234-1241. 2012.

[2] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu ,“3D convolution neural networks for human action recognition”, IEEE Transactions On Pattern Analysis And Machine Intelligence , vol. 35, no. 1,pp. 221-231, 2013.

[3] Meng Li, Howard Leung, and Hubert P. H. Shum,“Human action recognition via skeletal an depth based feature fusion,”Proceedings of the 9Th International Conference on Motion in Games,pp.123-132, 2016

[4] 1 Jeevan J.Deshmukh, 2Nita S.Patil, 3Dr.Sudhir D.Savarkar 1Student, 2 Assistant Professor, 3Professor of the IEEE, ISSN: 2321-9939.

[5] Yuanyuan Huang, Haomiao Yang,and Ping Huang, “Action recognition using hog features in different resolution video sequences,” International Conference on Computer Distributed Control and Intelligent Environmental Monitoring, 2012.

[6] Heng Wang, Alexander Klaser, Cordelia Schmid,and Cheng-Lin Liu,“Dense trajectories and motion boundary descriptors for action recognition,”International journal of computer vision,pp.60-79,2013.

[7] Jilin Communications Polytechnic, Changchun, pp.117-126.

[8] TamV.Nguyen et al.,“Spatial-Temporal Attention-Aware Pooling for Action Recognition,” IEEE Transactions On Circuits And Systems For Video Technology, vol. 25, no. 1,pp.77-86, 2015.

Thank you !