# Video Content Classification using Deep Learning

Prof. Saraswati Patil, Pradyumn Patil , Shruti Pisal , Yashraj Pawar and Vishwajeet Pawar

***Abstract*—video content classification is an important research content in computer vision, which is widely used in many fields, such as image and video retrieval, computer vision.This paper presents a model that is a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) which develops, trains and optimizes a deep learning network which can identify the type of video content and classify them into categories such as "Animation , Gaming ,natural content,flat content etc". To enhance performance of the model novel key-frame extraction method is included so as to classify only the key-frames,thereby by reducing the overall processing time without sacrificing any significant performance.**

***Index Terms*—Deep learning, Convolutional neural networks, Recurrent neural networks, image classification , Computer Vision.**

## I. Introduction

VIDEO has become more popular in many applications in recent years due to increased storage capacity, more advanced network architectures, as well as easy access to digital cameras, especially in mobile phones. Today people have access to a tremendous amount of video, both on television and the Internet. The amount of video that a viewer has to choose from is now so large that it is infeasible for a human to go through it all to find a video of interest. One method that viewers use to narrow their choices is to look for video within specific categories or genre. The sheer Amount of Video Data that is generated in today's age has motivated many researcher to find ways to classify the information in classes such as human Activities ,Research has begun on automatically classifying video.The excessive growth of video data generation as well as sharing through the Internet appeals for the efficient organization ,classification and analysis of videos. For this reason, it is necessary to have a system for generating relevant labels to a video or to different parts of video automatically without human intervention. The video content should be extracted and understood from the video data to find the type of video it belongs to.In this project we will create a classifier for video content.It will extract the key frames from a given video and will feed these frames to the CNN plus RNN classifier which will label it to a particular class i.e animation, news ,games,scenery etc .There is a recent trend to learn robust feature representations by using deep learning from video data. the two well known categories of deep learning algorithms for video classification, i.e., supervised deep learning and unsupervised feature learning.

## II. Literature Review

Over the last decade, active researchers have produced methods for improving Video Classification.. In [1],They proposed a Memory Efficient Model for Video Classification where they trained a computationally expensive teacher network which computes a representation for the video by processing all frames in the video.a relatively inexpensive student network whose objective is to process only a few frames of the video and produce a representation Then used the student network for classification thereby reducing the time required for processing the video The proposed approach in [2] used an improved CNN based classification algorithm with softmax loss function to classify mine video scenes .Compared to the older algorithms this CNN algorithm showed higher accuracy,precision and recall in classifying mine videos. It did solve the drawbacks that were there in the traditional model but as it requires massive data to give more accurate result which was not available at that moment so it was predicted that future massive data and complex training would make that model more efficient .The presented method in paper [3] considers the problem constructing a minimum representative set of video physical features that can be used in different regression and classification tasks when working with video. To test the resulting set of features, the problem of classifying video into 4 classes was considered: cartoon, drone shooting, computer game and sports broadcast. [4] uses support vector machine.Proposed System is capable of extracting key frames from videos and classifying them as natural scenery, personality, animal and plant.For feature extraction new methods like edge histogram , dominant color, color layout and face feature were used for image retrieval .Results have shown a high accuracy when parameters are C=14 and y =0.4.this research [5] paper proposes a novel method for sports video scene classification with the particular intention of video summarizing.In this paper discussed methodologies in the recent past for shot classification exclusively or part of their scheme, conclusively, deep-learning approaches presented high quality of the classification. Deep learning proved its importance in image classification, and incorporating the pre-trained model and applying augmentation . [6] In this paper, a template-based key-frame extraction method is proposed which employs action template-based similarity to extract key-frames for video classification tasks.Combining pre-trained CNN with ConvLSTM has achieved the highest classification accuracy among the other architectures.One of the limitations is that CNN architecture used did not produce the best results .Also it requires a machine with more than one GPU .So future work could be focused the on application of the proposed algorithm using more powerful architectures for real-world

video classification.

## III. Corresponding Concepts

### A. Deep networks

these the dominant approach for video understanding [ This includes the highly successful two-stream networks and 3D convolution networks . In this paper, we use 3D CNNs, but the long-term feature bank can be integrated with other families of video models as well

### B. Temporal and relationship models

include RNNs that model the evolution of video frames and multi-layer perceptrons that model ordered frame features. To model finer-grained interactions, a growing line of work leverages pre-computed object proposals or defections , and models their co-occurrence temporal order , or spatial arrangement [52] within a short clip.

### C. Spacio-temporal action localization

n is an active research area . Most recent approaches extend object detection frameworks to first propose tubelets/boxes in a short clip/frame, and then classify the tubelets/boxes into action classes . The detected tubelets/boxes can then be optionally linked to form full action tubes . In contrast to our method, these methods find actions within each frame or clip independently without exploiting long-term context

### D. Long-term video understanding

with modern CNNs is less explored, in part due to GPU memory constraints. One strategy to overcome these constraints is to use precomputed features without end-to-end training. These methods do not optimize features for a target task, and thus are likely sub-optimal. Another strategy is to use aggressive sub-sampling or large striding. TSN samples 3-7 frames per video. ST-ResNet uses a temporal stride of 15. To our knowledge, our approach is the first that enjoys the best of three worlds: end-to-end learning for strong short-term features with dense sampling and decoupled, flexible long-term modeling.

## IV. Approach Used in Video Classification

Because many videos are present in the real world, an effective way to classify those videos is important. The main aim of the video classification method is to classify whether the video is used for athletics, films, amusing videos, school, etc. There are three different ways of classifying video called audio, video and text [7]. Apart from these three methods we can also use hybrid approach (using one more method combined approach) to categories the videos. Figure 3 shows Taxonomy of Video Classification Approaches.

Most scholars used the approach as most knowledge dependent on the vision is interpreted by human beings. Some

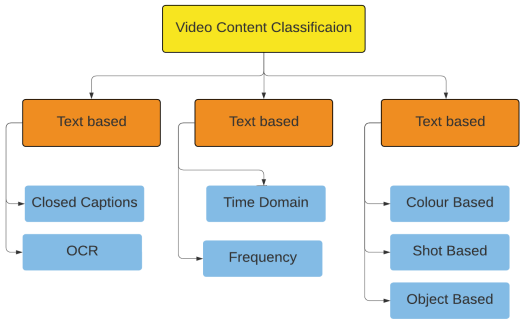



Fig. 1. Various Video Classification approach

authors have also where necessary coupled these visual aspects including audio and text [8]. Visual capability is primarily derived from image sequences or video files. Video's basic structure is like a combination of pictures is a fundamental part of video. Video may also be named as a collection of frames. Visual characteristics are typically dependent on colour, motion or shot time. These features must convey lighting, movement, background or video speed detail.

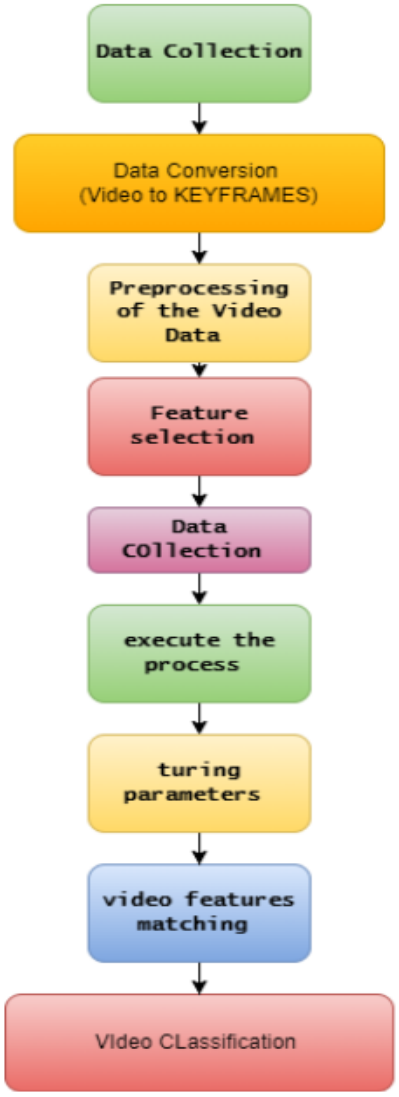
## V. Methodology

Video Classification technique involves some basic steps that have to be done in order to as show in the figure . To start the process we gather data from already available Data set like UCF [9] and Coin [10] . Both of contain mostly Video Data collected From YouTube. Then since we are dealing with frames of the video , Pre processing techniques are done such as key frame extraction and normalization. Then the main part of the process come that is classification where we extract features from video and classify the content and finally give a prediction,

A simple process flow diagram is a very general representation of the entire process and shows the various stages the video frames have to go through to get the intended result.

We will be further experimenting and working upon the pre -processing of video data i.e. key frame extraction and Neural Networks model that actually extract the features and compare them.

### A. Video Classification Model [8]




### B. Dataset

COIN data-set [10] consists of 11,827 videos related to 180 different tasks, which were all collected from YouTube. The average length of a video is 2.36 minutes. Each video is labelled with 3.91 step segments, where each segment lasts 14.91 seconds on average. In total, the data-set contains videos of 476 hours, with 46,354 annotated segments.

The UCF 50 [9] data set consists of realistic videos which are taken from YouTube. Our data set is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. For all the 50 categories, the videos are grouped into 25 groups, where each group consists of more than 4 action clips. The video clips in the same group may share some common features, such as the same person, similar background, similar viewpoint, and so on.

UCF50 [11] data set's 50 action categories collected from YouTube are: Baseball Pitch, Basketball Shooting, Bench Press, Biking, Biking, Billiards Shot,Breaststroke, Clean and Jerk, Diving, Drumming, Fencing, Golf Swing, Playing Guitar, High Jump, Horse Race, Horse Riding, Hula Hoop, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Lunges, Military Parade, Mixing Batter, Nun-chucks, Playing Piano, Pizza Tossing, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Skateboarding, Skiing, Skijet, Soccer Juggling, Swing, Playing Tabla, TaiChi, Tennis Swing, Trampoline Jumping, Playing Violin, Volleyball Spiking, Walking with a dog, and Yo Yo.

### C. Key-Frame Extraction

In video compression, a key frame, also known as an can be referred to a a frame in which a complete image is stored in the data stream.Since we know that in video This technique capitalizes on the fact that most video sources (such as a typical movie) have only small changes in the image from one frame to the next. Whenever a drastic change to the image occurs, such as when switching from one camera shot to another, or at a scene change, a key frame must be created. The entire image for the frame must be output when the visual difference between the two frames is so great that representing the new image incrementally from the previous frame would require more data than recreating the whole image.Key-frames are defined as the representative frames of a video stream, the frames that provide the most accurate and compact summary of the video content.

Frame extraction and selection criteria for key-frame extraction

Frame that are sufficiently different from previous ones using absolute differences in LUV color space Brightness score filtering of extracted frames [12] Entropy/contrast score filtering of extracted frames K-Means Clustering of frames using image histogram Selection of best frame from clusters based on and variance of laplacian (image blur detection)

### D. Neural Network models

**Algorithm 1** Classifiers

**Number of Classes = n**
**Number of Frames = M**
**Number of Videos = S**
$classes\_C \leftarrow |\{\mathbf{c}_1, \mathbf{c}_2.\mathbf{c}_3....\mathbf{c}_n\}|$

For each unique 'n' number of classes
we will be predicting the probability 'P' for frames 'F' :-

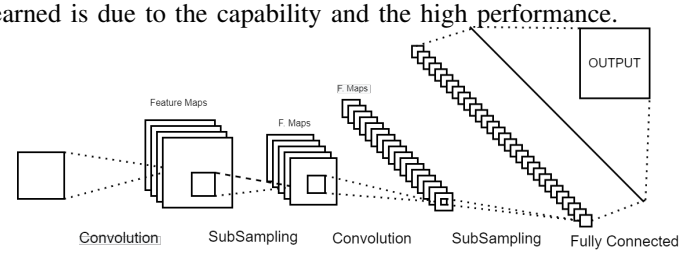
$$P(\mathbf{C}_n|S) = (\kappa(\mathbf{F}_0 + \mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_{M-1}))$$

The mentioned "k" can be a Neural Network which does the job of predicting the frames 'F'. [1] Now we discuss in detail about the model architecture we are going to implement this neural network.

### E. CNN architecture (spacial)

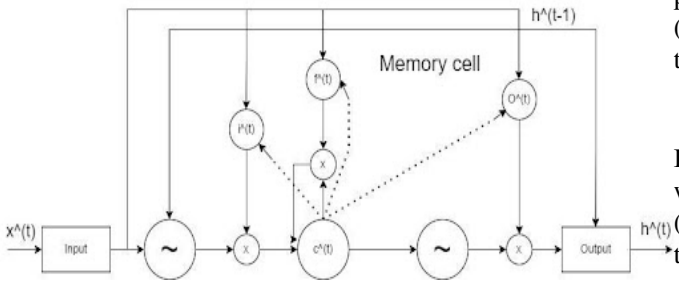
Neural networks are considered a very powerful and intelligent technique to classify different types of data for important applications. [13] [14] New types of neural networks called convolutional neural networks (ConvNets or CNNs).were developed classification and recognition tasks for many applications. It makes a breakthrough in image processing, video,

speech, and text recognition research, and the concentration it earned is due to the capability and the high performance.




*F. LSTM architecture (temporal)*

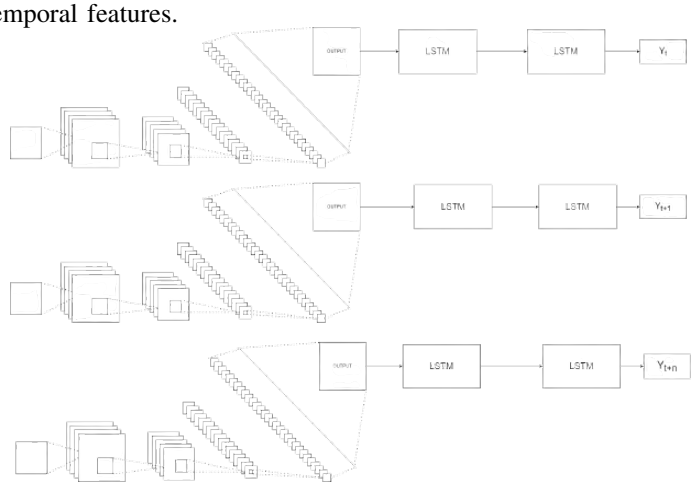
This is an RNN variant that was designed to store and access information in a long time sequence and makes perfect sense when used in . Unlike standard RNNs, nonlinear multiplicative gates and a memory cell are introduced. These gates, including input, output, and forget gates, govern the information flow into and out of the memory cell. The structure of an LSTM unit is illustrated in




These explicitly designed memory units and gates, LSTM is able to exploit the long-range temporal memory and avoids the issues of vanishing/exploding gradients [14] [15].

## VI. SpacioTemporal Neural Networks(CNN+RNN)

As we are working on videos we need to look at both spatial as well as temporal factor to make sure system understands the coso we need a hybrid model for classification of video.And we know that CNN i.e. Convolutional neural networks can take in an input image, assign importance to various aspects/objects in the image and be able to differentiate one from the other which means it can work on the spatial features of the video while RNN i.e Recurrent Neural networks which uses sequential data or time series data so it can work on the temporal features.




**Models Layer's** :-

https://sci-hub.se/10.1109/iacs.2018.8355453 It contains a set of digital filters to perform the function of convolution in the input data.It is the number of filters the convolutional layers will learn from. It is the absolute value and determines the number of filters that come out of convolution. It uses the 'Relu'activation function here.

**Average Pooling Layer**:-
It involves calculating the average for each part of the feature map. This means that each 2×2 square of the feature map is down sampled to the average value in the square.

**Max Pooling Layer** :-
It is an operation that calculates the maximum value for patches of a feature map, and uses it to create a downsampled (pooled) feature map.It is used to reduce the dimensions of the feature maps.

**DropOut Layer**:-
It is a technique used to prevent a model from overfitting.It works by randomly setting the outgoing edges of hidden units (neurons that make up hidden layers) to 0 at each update of the training phase.

**Dense Layer**:-
We can call it a deeply connected network. Each neuron in the dense layer receives input from all neurons of its previous layer.Acts like output layer and givas results according to classified classes.

**SoftMax Layer**:-
It assigns decimal probabilities to each class in a multi-class problem. Those decimal probabilities must add up to 1.0. This additional constraint helps training converge more quickly than it otherwise would.It is implemented through a neural network layer just before the output layer.

*A. WorkFlow*

- Firstly the dataset is loaded
- The images frames are resized to a common scale of 224 x224.
- As we do not have any separate testing dataset we will be dividing the current dataset into training and validation dataset. in the ratio of 80:20
- The datasets are loaded with a batch size of 128 and then the dataset is shuffled to avoid biased conditions.the class mode will be categorical.
- Then the images are preprocessed for training using feature extraction and edge detection methods.
- The above process gives a compiled model so from here the model has to be fit to train. So the model is trained over the training dataset under 15 to 30 epochs.
- For the Testing Part ,Video input is taken and the key frames are then feeded in the pretrained model.where the input frame image of size 224 x 224 is given to the model.

- The result of prediction is given in the form of percentage where it tells by what percentage the video resembles each class.

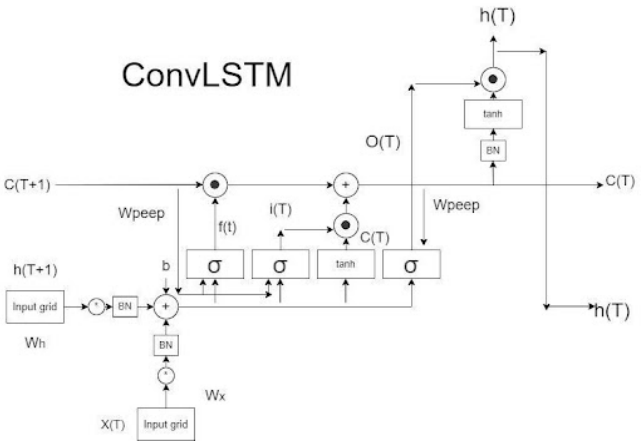
## VII. Variaous Approch for Video Classification

### A. ConvLSTm Approach

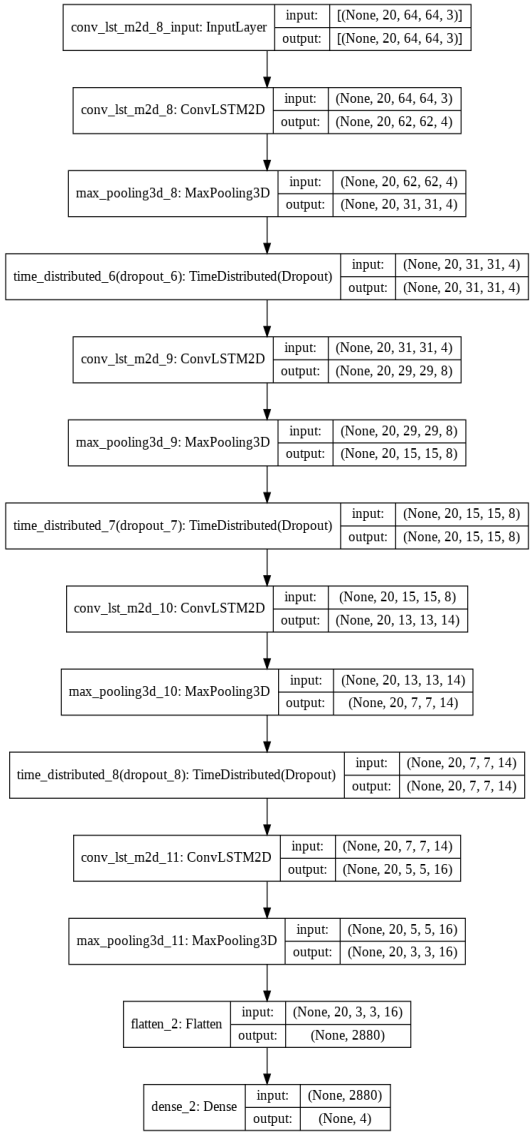
We extend the idea of FC-LSTM [16] to ConvLSTM which has convolutional structures in both the input-to-state and state-to-state transitions. this can be done by By stacking multiple ConvLSTM layers and forming an encoding-forecasting structure [17] [18], we can build an end-to-end trainable model for our classification purpose.

if we assume a spatial region represented by an M × N grid which consists of M rows and N columns. each cell containing P measurements which vary over time since this is a temporal base classifier and videos changes their content over time.. Thus, the observation at any time can be represented by a tensor X RP ×M×N , where R denotes the domain of the observed features. If we record the observations periodically, we will get a sequence of tensors Xˆ 1, Xˆ 2, . . . , Xˆ t. The spatiotemporal sequence forecasting problem is to predict the most likely length-K sequence in the future given the previous J observations which include the current one: Xˇ t+1, . . . , Xˇ t+K = arg max Xt+1,...,Xt+K p(Xt+1, . . . , Xt+K — Xˆ tJ+1, Xˆ tJ+2, . . . , Xˆ t) (1) For precipitation nowcasting, the observation at every timestamp is a 2D radar echo map. If we divide the map into tiled non-overlapping patches and view the pixels inside a patch as its measurements (see Fig. 1), the nowcasting problem naturally becomes a spatiotemporal sequence forecasting problem

ConvLSTM cell is a variant of an LSTM network that contains convolutions operations in the network [15]. It is an LSTM with convolution embedded in the architecture, which makes it capable of identifying spatial features of the data while keeping into account the temporal relation.




For video classification, this approach effectively captures the spatial relation in the individual frames and the temporal relation across the different frames. As a result of this convolution structure, the ConvLSTM is capable of taking in 3-dimensional input (width, height, num of channels) whereas a simple LSTM only takes in 1-dimensional input hence an LSTM is incompatible for modeling Spatio-temporal data on its own.




### B. LRCN approach

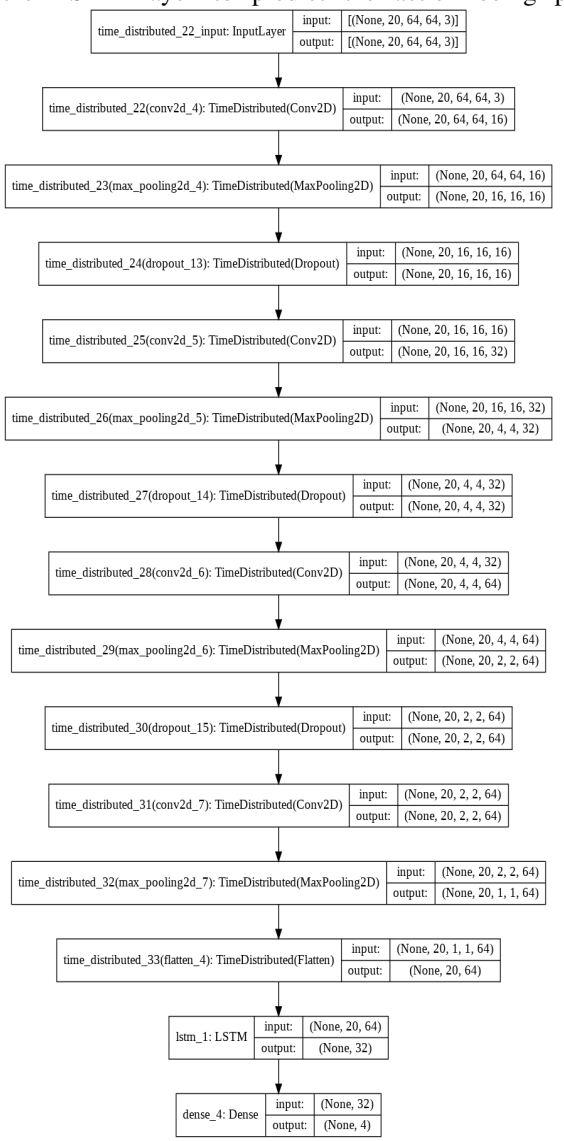
Implement the LRCN Approach by combining Convolution and LSTM layers in a single model [15]. Another similar approach can be to use a CNN model and LSTM model trained separately. The CNN model can be used to extract spatial features from the frames in the video, and for this purpose, a pre-trained model can be used, that can be fine-tuned for the problem. And the LSTM model can then use the features extracted by CNN, to predict the action being performed in the video.

But here, we will implement another approach known as the Long-term Recurrent Convolutional Network (LRCN), which combines CNN and LSTM layers in a single model. The Convolutional layers are used for spatial feature extraction from the frames, and the extracted spatial features are fed to LSTM layer(s) at each time-steps for temporal sequence modeling. This way the network learns spatiotemporal features directly in an end-to-end training, resulting in a robust model.

We will also use TimeDistributed wrapper layer, which allows applying the same layer to every frame of the video in-

dependently. So it makes a layer (around which it is wrapped) capable of taking input of shape (no_of_frames, width, height, num_of_channels) if originally the layer's input shape was (width, height, num_of_channels) which is very beneficial as it allows to input the whole video into the model in a single shot.

To implement our LRCN architecture, we will use time-distributed Conv2D layers which will be followed by Max-Pooling2D and Dropout layers. 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.

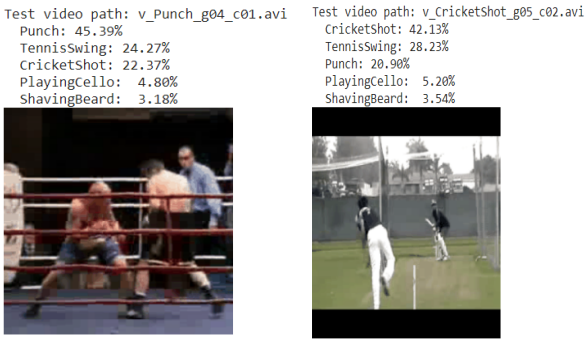



## VIII. Results and Discussion

### A. Inception V3 model

The effects of key-frame extraction on Inception V3 provided by Keras-Applications [19].

The model has attained greater than 78.1 accuracy in about 170 epochs on each of these according to google [20] The model after compilation is trained and then tested including Key-frame extraction , the model has given accuracy of 77.23This accuracy was possible because of introducing some complex layers like Dense layer and Dropout layer where the random neurons are dropped for training.Also as we used soft-max function which helped in quicker training coverage concluding that Key-frames extraction maintained same levels of accuracy with faster processing time supporting our assumption .

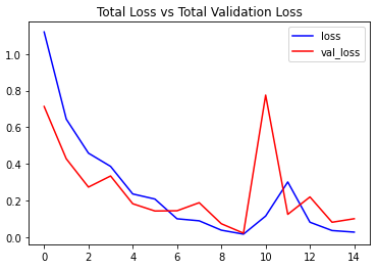


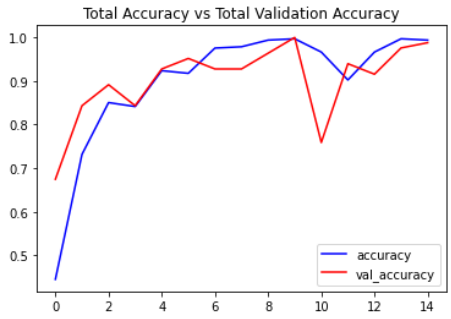

### B. ConvLSTM results

Graph of Models and Accuracy curves

| no.of epochs | loss | Accuracy | val loss | val accuracy |
|---|---|---|---|---|
| 1 | 1.3917 | 0.2774 | 1.3715 | 0.3562 |
| 2 | 1.3410 | 0.3801 | 1.2734 | 0.5204 |
| 3 | 1.1869 | 0.5068 | 1.2266 | 0.4110 |
| 4 | 1.0063 | 0.5685 | 1.0510 | 0.4932 |
| 5 | 0.8963 | 0.6096 | 0.7548 | 0.6301 |
| 6 | 0.7132 | 0.6712 | 0.7022 | 0.7123 |
| 7 | 0.5275 | 0.7979 | 0.4703 | 0.8356 |
| 8 | 0.4196 | 0.8493 | 0.5588 | 0.9277 |
| 9 | 0.3807 | 0.8459 | 0.5261 | 0.9639 |
| 10 | 0.3322 | 0.8733 | 0.4099 | 1.000 |
| 11 | 0.1154 | 0.9349 | 0.7755 | 0.7590 |
| 12 | 0.3015 | 0.9024 | 0.1249 | 0.9398 |
| 13 | 0.0818 | 0.9665 | 0.2198 | 0.9157 |
| 14 | 0.0364 | 0.9970 | 0.0817 | 0.9759 |
| 15 | 0.0278 | 0.9939 | 0.1009 | 0.9880 |

evaluation accuracy = 0.9708



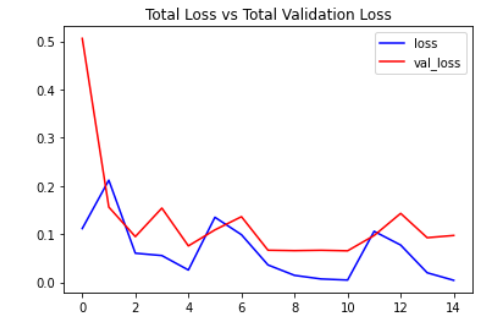


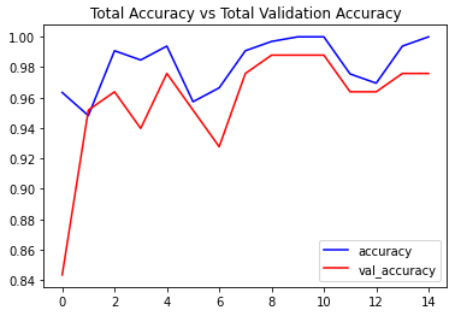

### C. LRCN

Graph of Models Loss and Accuracy Curves

| no.of epochs | loss | Accuracy | val loss | val accuracy |
|---|---|---|---|---|
| 1 | 1.3965 | 0.2979 | 1.3717 | 0.1384 |
| 2 | 1.3252 | 0.3425 | 1.4161 | 0.1781 |
| 3 | 1.2344 | 0.4555 | 1.3471 | 0.2877 |
| 4 | 1.1417 | 0.5240 | 0.9880 | 0.5753 |
| 5 | 1.0090 | 0.5856 | 0.9365 | 0.6164 |
| 6 | 0.9444 | 0.5959 | 0.9169 | 0.6575 |
| 7 | 0.8739 | 0.6233 | 0.7460 | 0.7260 |
| 8 | 0.7811 | 0.6781 | 0.8978 | 0.6438 |
| 9 | 0.7387 | 0.6986 | 0.6321 | 0.7534 |
| 10 | 0.6822 | 0.7260 | 0.5543 | 0.8219 |
| 11 | 0.2312 | 0.9116 | 1.2060 | 0.8493 |
| 12 | 0.0357 | 0.9848 | 0.1181 | 0.7945 |
| 13 | 0.0053 | 1.000 | 0.0418 | 0.8493 |
| 14 | 0.0039 | 1.000 | 0.0382 | 0.9178 |
| 15 | 0.0020 | 1.000 | 0.0449 | 0.9880 |

evaluation accuracy = 0.9635




### D. Testing on Random Videos

| Inception V3 Single Video Prediction | | | |
|---|---|---|---|
| Video ID | Type of Content | predicted Video | Confidence |
| guitar string | guitar playing | guitar playing | 98.8385 |
| Parks View | Swing | Swing | 97.6374 |
| Walking Dog in Snow | Walking with Dog | Walking with Dog | 86.0622 |
| swing Chair in porch | Swing | Walking with Dog | 58.1278 |
| Army day Parade | Military Parade | Military Parade | 98.2633 |
| Horse Racing Track | Horse Race | Horse Race | 83.3013 |
| Wedding piano | Playing piano | Playing piano | 99.9551 |

| LRCN Single Video Prediction | | | |
|---|---|---|---|
| Video ID | Type of Content | predicted Video | Confidence |
| guitar string | guitar playing | guitar playing | 98.8385 |
| Parks View | Swing | Swing | 97.6374 |
| Walking Dog in Snow | Walking with Dog | Walking with Dog | 86.0622 |
| swing Chair in porch | Swing | Walking with Dog | 58.1278 |
| Army day Parade | Military Parade | Military Parade | 98.2633 |
| Horse Racing Track | Horse Race | Horse Race | 83.3013 |
| Wedding piano | Playing piano | Playing piano | 99.9551 |

| ConvLSTM Single Video Prediction | | | |
|---|---|---|---|
| Video ID | Type of Content | predicted Video | Confidence |
| guitar string | guitar playing | guitar playing | 96.7585 |
| Parks View | Swing | Swing | 98.65164 |
| Walking Dog in Snow | Walking with Dog | Walking with Dog | 76.0923 |
| swing Chair in porch | Swing | Walking with Dog | 62.1458 |
| Army day Parade | Military Parade | Military Parade | 88.3443 |
| Horse Racing Track | Horse Race | Horse Race | 81.3241 |
| Wedding piano | Playing piano | Playing piano | 89.2352 |

TABLE I
PERFORMANCE ANALYSIS.

| Model | UCF 50[1] | | | UCF 101 | | |
|---|---|---|---|---|---|---|
| | Training Accuracy | evaluation Accuracy | Testing Accuracy | Training Accuracy | evaluation Accuracy | Testing Accuracy |
| Inception v3 | - | - | - | - | - | 72.23 |
| LRCN | 97.98 | 93.81 | 96.35 | 97.83 | 93.83 | 88.43 |
| ConvLSTM | 96.61 | 88.71 | 97.08 | 97.98 | 93.81 | 96.35 |

Note: that the classifiers used are same in all the cases
Number of Epochs = 15

| Model Type (x) | Model type (y) | Mean Difference | SE | p value | Confidence interval |
|---|---|---|---|---|---|
| LRCN | LRCN(1) | -0.004 | 0.003 | 0.612 | |
| | ConvLSTM | -0.016 | 0.003 | 0.000 | |
| | ConvLSTM(1) | -0.035 | 0.003 | 0.000 | |
| LRCN(1) | LRCN | 0,004 | 0.003 | 0.612 | |
| | ConvLSTM | -0.012 | 0.003 | 0.000 | |
| | ConvLSTM(1) | -0.031 | 0.003 | 0.000 | |
| ConvLSTM | LRCN | 0.016 | 0.003 | 0.000 | |
| | LRCN(1) | 0,012 | 0.003 | 0.000 | |
| | ConvLSTM(1) | -0.020 | 0.003 | 0.000 | |
| ConvLSTM(1) | LRCN | -0.035 | 0.003 | 0.000 | |
| | LRCN(1) | 0.031 | 0.003 | 0.000 | |
| | ConvLSTM | 0.020 | 0.003 | 0.000 | |

TABLE II
COMPARISONS USING TUKEY'S HSD ON YOUTUBE VIDEOS DATASET

Note: Calculations based on TABLE 1 results.
(1) denotes use of key-frame extraction
(1) value significanceconsidered at 0.05 levels

### *E. Comparision*

## IX. CONCLUSION

We have gone through a couple of deep learning techniques key topics related to video analysis, that is video classification. We mostly rely on the modeling of the spatial and temporal information in videos. we showcased the , the essence of deep learning for video classification is to derive robust and discriminatory feature representations from raw data through exploiting massive videos with an aim to achieve effective and efficient recognition, which could hence serve as a fundamental component in multiple video analysis tools, ConvLSTM and LRCN both of these spacio-temporal neural networks give promising result suggests that LRCN was more efficient when it came to resource consumption with comparison to other models and maintained sma levels of accuracy proving the conclusions done in previous research [15]

Though extensive efforts have already been made in video classification with deep learning, we believe we are just beginning to discover the potential of deep learning in the era where Big DATA is growing exponentially . Given the substantial amounts of videos generated at an astounding speed every hour and every day, it remains a challenging open problem how to derive better video representations with deep learning modeling the abundant interactions of objects and their evolution over time with limited (or without any) supervisory signals to facilitate video content understanding (i.e., the recognition of human activities and events as well as the generation of free-form and open-vocabulary sentences for describing videos). We hope this chapter sheds light on the nuts and bolts of video classification and captioning for both current and new researchers.

## X. LIMITATIONS

- The Processing was carried out on low-end hardware resulting in less processing power therefore a small sub sample of Data set was used to conduct the tasks.
- Absence of a dedicated 'GPU' makes the computational task a bit time consuming.
- At a time only 4 classifiers were trained , any more than that may affect the result as many classifier have inherently very similar features.

## XI. FUTURE SCOPE AND APPLICATIONS

- Use the model for more specific purposes related to 'video classification' such as 'Action Recognition' ,

'Video Caption generator ' [14] 'Event Detection in a video 'and etc.

- Building Vision Systems to understand adverts [2] and can work in conjunction with audio processing analysis.
- Video Intelligent systems involving object tracking.
- Use container-orchestration system Tools such as 'Docker' and 'Kubernetes' for easy deployability and scalability .
- Integrate Cloud processing platforms such as 'Google cloud' so increase the efficiency of the overall system . [21]

## REFERENCES


[1] S. Bhardwaj, M. Srinivasan, and M. M. Khapra, "Efficient video classification using fewer frames," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 354–363.

[2] O. Ye, Y. Li, G. Li, Z. Li, T. Gao, and T. Ma, "Video scene classification with complex background algorithm based on improved cnns," in *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. IEEE, 2018, pp. 1–5.

[3] R. Kazantsev, S. Zvezdakov, and D. Vatolin, "Application of physical video features in classification problem," *International Journal of Open Information Technologies*, vol. 7, no. 5, pp. 33–38, 2019.

[4] Z. Min, "Scenery video type classification based on svm," in *2009 Asia Pacific Conference on Postgraduate Research in Microelectronics & Electronics (PrimeAsia)*. IEEE, 2009, pp. 287–289.

[5] M. Rafiq, G. Rafiq, R. Agyeman, G. S. Choi, and S.-I. Jin, "Scene classification for sports video summarization using transfer learning," *Sensors*, vol. 20, no. 6, p. 1702, 2020.

[6] R. Savran Kızıltepe, J. Q. Gan, and J. J. Escobar, "A novel keyframe extraction method for video classification using deep neural networks," *Neural Computing and Applications*, pp. 1–12, 2021.

[7] Z. Kastrati, A. S. Imran, and A. Kurti, "Integrating word embeddings and document topics with deep learning in a video classification framework," *Pattern Recognition Letters*, vol. 128, pp. 85–92, 2019.

[8] M. Islam, S. Sultana, U. Roy, and J. Al, "A review on video classification with methods, findings, performance, challenges, limitations and future work," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. Vol 6, No 2 (2020), pp. 47–57, 12 2020.

[9] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[10] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "Coin: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1207–1216.

[11] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine vision and applications*, vol. 24, no. 5, pp. 971–981, 2013.

[12] Keplerlab, "Katna python," https://github.com/keplerlab/katna, 2021.

[13] I. Aljarrah and D. Mohammad, "Video content analysis using convolutional neural networks," in *2018 9th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2018, pp. 122–126.

[14] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," in *Frontiers of multimedia research*, 2017, pp. 3–29.

[15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[16] J. Sun, M. Xue, J. W. Wilson, I. Zawadzki, S. P. Ballard, J. Onvlee-Hooimeyer, P. Joe, D. M. Barker, P.-W. Li, B. Golding *et al.*, "Use of nwp for nowcasting convective precipitation: Recent progress and challenges," *Bulletin of the American Meteorological Society*, vol. 95, no. 3, pp. 409–426, 2014.

[17] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[20] g. c. cloud TPU, "Advanced guide to inception v3," in *Google. [Online]. Available*.

[21] E. Bisong, *Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners*. Apress, 2019.