# Video-based Action Recognition Using Deep Learning

A PROJECT REPORT

*Submitted by*

**ANUSHREE (TKM20MCA-2010)**

to

The APJ Abdul Kalam Technological University

*In partial fulfillment of the requirements for the award of the degree of*

**MASTER OF COMPUTER APPLICATIONS**


Thangal Kunju Musaliar College of Engineering
Kerala

**JULY 2022**

# DECLARATION

I hereby declare that the project report entitled "**Video-based Action Recognition Using Deep Learning**", which I submitted in partial fulfillment of the requirements for the award of degree of Master of Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a genuine work produced by me under the guidance of Prof. JASMIN M R. In my contribution, I have used my own thoughts to describe my ideas, and where I have borrowed ideas or words from others, I have properly and fully referenced and recognised those authors as my sources. I further vouch that I have adhered to the standards of academic integrity and honesty and have not fabricated or distorted any facts in my contribution I am aware that any breach of the aforementioned rules might result in disciplinary action by the institution, the university, and/or the sources who were improperly referenced or from whose sufficient permission was not sought. This report has never before used as the foundation for the award of a degree, diploma, or other comparable title by another university.

KOLLAM ANUSHREE
18/07/2022

# DEPARTMENT OF COMPUTER APPLICATIONS

# TKM COLLEGE OF ENGINEERING


## C E R T I F I C A T E

This is to certify that, the project report entitled **"VIDEO-BASED ACTION RECOGNITION USING DEEP LEARNING"** is submitted by **ANUSHREE (TKM20MCA-2010)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the degree of Master of Computer Applications, is a bonafide record of the project work carried out by her under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor Head of the Department External Examiner

# ACKNOWLEDGEMENT

For the accomplishment of this undertaking, I am very grateful to God the Almighty and our parents. I owe everyone my deepest gratitude and sincere thanks for lending their invaluable time and knowledge to the successful conclusion of my project.

I owe a huge debt of gratitude to **Dr. FOUSIA M SHAMSUDEEN**, Assistant Professor and Head of the Department, MCA, TKMCE, for her unwavering support and inspiration during the project's development.

I want to express my heartfelt appreciation to my project's advisor, Prof. JASMIN M. R., Assistant Professor, Department of Computer Applications, TKMCE, for her professional direction, cooperation, and tremendous encouragement throughout the project.

I also want to convey my thanks to the whole teachers and staff at the TKMCE Department of Computer Applications, who have supported me throughout my academic career.

I would also want to appreciate my family and friends for their encouragement and support in seeing this project through to its final conclusion.

**ANUSHREE**

# ABSTRACT

One of the key detection methods that offers benefits in a variety of fields, including video surveillance, video captioning, security, content censorship, and military applications, is human action recognition. The main goal of this suggested research is to efficiently categorise many actions from a video in order to create the action label that each action represents. Convolution neural networks (CNNs) are a subset of deep learning models that can operate directly on raw data; as a result, we may utilise this subset of models to extract the appropriate relevant features from the data and transform them to generate a series of frames with comparable actions. Later, these frames can be used as a sequential input to feed the LSTM network for action prediction. In this effort, a network termed the LRCN (Long-Term Recurrent Convolutional Network) was created by combining these two networks. In order to accomplish the project's objective, various networks are leveraged, and their performance is assessed. Numerous publicly accessible datasets, including UCF50, are used to assess the paper. These datasets' results indicate that the suggested approach produces superior results in terms of overall accuracy.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| Abbreviation | Definition |
|---|---|
| HAR | Human Action Recognition |
| CNN | Convolutional Neural Network |
| LSTM | Long Short-Term Memory |
| RNN | Recurrent Neural Network |
| LRCN | Long-Term Recurrent Convolutional Network |

# Chapter 1
# Introduction

Since actions may be performed by many people in varied settings, human action recognition (HAR) is still a difficult problem. Various deep learning techniques have enhanced the performance of video action recognition. These days, more scholars are interested in this topic of study because of all the potential applications. Applications of HAR include the recognition of patient behaviour in hospitals, the detection of anomalous activity in sensitive locations, sports data analysis, video retrieval, etc. Although there are differences within each activity class and parallels between action classes, the HAR paradigm has difficulties in effectively identifying the activities. Inter-class similarity occurs when behaviours like "run," "jog," and "walk" are recognised because of how similar their shapes of actions are. Variations within a class action occur from various people carrying out the same action in different ways. Similarly, the camera perspective affects how an activity seems.

Simple hand and leg motions are what define human activity. In complicated situations, it combines movement of the body with engagement with the environment. Therefore, only one image cannot accurately convey an activity. It involves discovering the connections between the data taken from each frame. Recognizing the sort of action in an action movie depends on the time link. This project's main objective is to recognise activities in a given video and then forecast the behaviours that the person in the video will take. Since there may be fluctuation, this may be quite difficult. As the duration of the time period for the same and other sorts of acts may vary, this can be quite difficult. The usage of these deep learning techniques has been discovered in a variety of domains, including virtual reality, military applications and security as well as surveillance and health systems. In recent times, researchers have adopted LSTMs based on sequential learning for HAR. An LSTM's component pieces are hidden units, input units, and output units. The LSTM's primary competitive advantage is the ability of its memory cells to remember earlier data or learning parameters.

In the field of computer vision (CV), deep learning has lately shown positive outcomes. Deep learning creates models by imitating the way that the human brain processes information. Various processing layers, including convolutional layers, fully connected layers, MaxPooling, ReLu, and Softmax function, are used to create these models. You may use the CNN model to analyse multidimensional data. There are several publicly accessible datasets for action recognition, including UCF50, UCF101, HMDB-51, KTH, etc. These datasets contain a variety of actions, including diving, running, horseback riding, pushups, kicking, boxing, throwing, swinging, playing Guitar, and many others.

Other significant HAR issues include the following: Because of complex backgrounds, imbalanced datasets that affect a CNN model's learning, and the fact that background complexity, lighting conditions, shadows and extract irrelevant information when using traditional methods of classifying human action, it is difficult to identify the right human activities from real-time observations. This leads to ineffective action classification. An enormous amount of training samples are required for a CNN model in order to learn, and feature extraction from the full video sequence would provide a number of irrelevant characteristics that would lower classification accuracy.

Because LRCN models are simpler to tune and can gain accuracy from much greater depth, I employed them to obtain improved accuracy in the proposed study to solve these current obstacles. This will thus address the overfitting issues.

## 1.1 Problem Statement

Many different industries use human action recognition (HAR), including video surveillance, identity verification, and the creation of smart systems for human-machine interaction. It is a difficult topic in the realm of machine learning and computer vision. The ability to recognize human behavior depends on a strong feature representation. For HAR, it is not enough to just detect characteristics and focus on a certain spatial area; we must also think about how features have evolved through time.

Recent years have seen the introduction of many action representation techniques, including key point tracking trajectory features, regional and global features depending on temporal and geographical changes, motion changes based on depth information, and depending on alterations in a person's position. Using the effective use of deep learning for object recognition and picture categorization, numerous. Deep learning has also been used by researchers to recognize human activity.

Due to the problem of computation intensity and inefficiency in correctly classifying the actions, a combination of CNN and RNN i.e., LRCN (Long Term Recurrent Convolutional Network) model is used.

## 1.1 Objective

This project's main goal is to employ deep learning techniques to recognise different activities in videos. The major goal of this research is to use the model to precisely identify distinct activities in a collection of video frames. The LRCN (Long Term Recurrent Convolutional Network) model, which is a combination of CNN and RNN, is used to accomplish the following:

- handle overfitting issues
- speed up the training process
- achieve better accuracy

# Chapter 2

# Related Work

Two categories of action recognition techniques are now in use: I those that employ manually created visual characteristics and (ii) those that use deep learning. A priori model must be taught about the properties of the images in order to use the hand-crafted picture feature-based techniques. Contrarily, deep learning no mathematical representation of the characteristics is necessary; Instead, during training, it finds the fundamental traits. The foundation of conventional action recognition techniques is various feature detectors, descriptors, and feature trajectories. Different feature descriptors have been used in the past.

Nilay Tufek presented a method to develop an action recognition system using only a little quantity of accelerometer and gyroscope data [1]. A variety of products are created and put into use, such as a wireless radio frequency module called ZigBee, an accelerometer, and a gyroscope. They also have sensors for humidity, temperature, and heart rate. These sensors were used in this study to capture data at a frequency of 5 Hz while individuals engaged in 7 distinct activities. The dataset is known as the ETEXWELD dataset since it was created through experimental research conducted as part of a European Union project (ETEXWELD H2020: RISE. 644268). Several experiments using gyroscopes and accelerometers to identify activity and authorisation have been published in the literature. Also, it is shown how to use data augmentation and data balancing approaches to enhance accuracy, precision, recall, and f1-score metrics. It has achieved significantly greater accuracy and precision after using data set balancing. After balancing, data augmentation is applied, and it also helped the outcomes. Only the training set is used for all these treatments. To test how to model optimise on training datasets, 10 fold cross validation was utilised throughout the construction of all procedures.

With a 97.4% accuracy rate, this study's 3 layer LSTM model produced a brand-new state-of-the-art result on the UCI HAR dataset. Additionally, the same model was applied to the gathered dataset (ETEXWELD), and a 99.0% accuracy rate was found, indicating a significant contribution. Additionally, the

performance evaluation takes into account measures for precision, f1-score and recall in addition to accuracy findings. A three-layer LSTM network was also used to develop a real-time application, which was used to examine how effectively the best model could categorise actions.

Since the existence of the position and temporal fluctuations in the action video makes the recognition difficult, S. P. Sahoo, S. Ari, and K. Mahapatra [2] have presented a method for action recognition. In this study, the innovative idea of a depth history picture is combined with sequence and shape learning to overcome these issues (DHI). For sequential learning, a deep bidirectional long short-term memory (DBiLSTM) is built to describe the temporal link between the action frames. A trained convolutional neural network is utilised to extract the action information from each frame (CNN). The DHI is developed by estimating and projecting the depth information from each action frame onto the X-Y plane. From a set of DHI pictures, the fine-tuned network is utilised to identify activities. To significantly enhance the training set and avoid the network becoming overfit, data augmentation is performed. Performance accuracy for the proposed work is 97.67%, 95.00%, 73.13%, 92.97%, and 69.74%, respectively, when tested against publically available datasets including KTH, UCF athletics, JHMDB, UCF101, and HMDB51.

Khan et al.[3] recommended using a deep learning-based system to categorise and forecast video tags. For the language used in this study, the authors created a dataset and tag vocabulary. They also suggested a powerful shot identification technique to locate movie keyframes. However, their approach disregards the video's motion data, which is more relevant to take into consideration for action recognition. To mimic the conventional 3D convolution, Yang et al. [4] developed asymmetric, unidirectional 3D convolutions. The asymmetric 3D convolution performs better than conventional 3D-CNN models in terms of efficacy and efficiency while using fewer parameters and a fraction of the processing resources.

For the image classification issue, Krizhevsky et al. [5] made the crucial advancement in deep learning. The objective was to train a model using an efficient method on 1.2 million high-resolution pictures so that it could categorise fresh images into 1000 distinct categories. This method used a deep convolutional neural network as its foundation. Later, a 2D pre-trained convolution network was utilised by Karpathy et al. [6] to investigate the notion of combining the temporal data from successive video frames. Due to the dataset's insufficient variety, the trained Spatio-temporal features failed to capture motion characteristics. Learning intricate features proved to be challenging, and the

results were noticeably worse than those obtained using hand-crafted feature-based algorithms.

In Tran et al. [7] research, he has employed 3D convolutions at the video level. The Sports1M1 dataset was used to train a 3D convolution-based network for feature extraction. It was computationally expensive to train such massive networks, and long-range temporal modelling remained difficult. To mimic the conventional 3D convolution, Yang et al. [8] developed asymmetric, unidirectional 3D convolutions. The asymmetric 3D convolution performs better than conventional 3D-CNN models in terms of efficacy and efficiency while using fewer parameters and a fraction of the processing resources.

According to Sargano et al., a pre-trained deep CNN architecture and SVM-KNN hybrid classifier may be used to classify human activity[9]. They performed more accurately than manually developed feature-based methods on the UCF sports and KTH datasets. Rafiq et al. published a method for categorising sports video situations utilising a pre-trained AlexNet convolutional neural network. concentrating on video summarization [10]. They assessed their outcomes using cricket films and contrasted their methodology with several deep-learning models. The Histogram of Oriented Gradients (HOGs) of the frames in each shot was used by Elharrouss et al. [11] to present a technique for multiple human activity detection, identification, and summarization in films. In order to determine the activities, this approach compares the created HOG to the pre-existing HOGs during the training phase.

The literature study makes clear that action recognition has been the subject of a wealth of research. CNNs were originally used for action detection on a frame-by-frame basis, and it was discovered that they outperformed manually created features-based methods. Later, by processing several frames simultaneously, 3D CNNs fared significantly better. Recently, CNNs and RNNs have been integrated in certain architectures to incorporate motion data. However, the performance of the current methods is not encouraging. By combining a CNN with an LSTM and analysing just the crucial frames of a video to minimise computational complexity and eliminate redundancy, the action recognition approach proposed in this research addresses the inadequacies of the prior methods.

# Chapter 3

# Methodology

## 3.1 PROPOSED SYSTEM

The basic goal of this project is to develop a system that, given a video, can identify the activity being performed in the movie and classify it. To do this, a model is developed that can sort the videos into the appropriate categories.

A sizable amount of data was needed to develop the model, which was then utilised to offer a solution to the categorization difficulty mentioned before. The UCF50 dataset, which is covered in the part after, was utilised in this study. To execute the apps, "Google Colab" was also utilised. Due of the roughly 12 GB of RAM and changeable GPU it offers, depending on the volume of traffic.

The project attempts to pursue the following goals:

1. The input video will be segmented into sets of frames, after which the human body part will be recognised in a series of frames,
2. After that, CNN will be used to identify the human activity, and deep learning techniques will be used to forecast the action frame.

## 3.2 BLOCK DIAGRAM

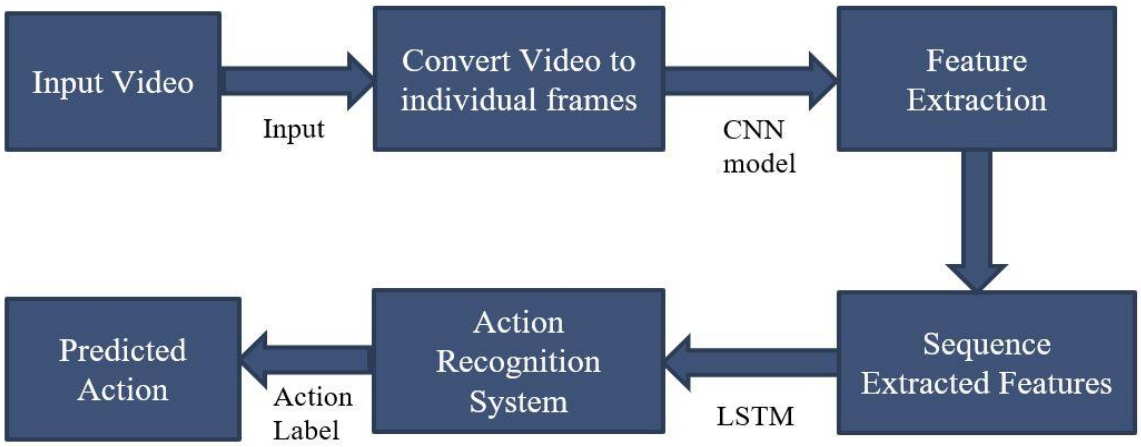


Figure 1: Block diagram of the system

The above figure depicts the block diagram for the proposed system, First the input video will be provided to the system and it takes this video and convert the video into a different set of frames after which they are resized to a suitable size. Following that, features are extracted for each action category by the CNN model, which are then progressively provided to the LSTM network for individual frame identification. As a result, it generates all action labels with regard to the category to which it belongs. This is accomplished using the LRCN model, which is covered in more detail in the section below.

## 3.3 DATASET

The dataset utilised in this research is an Action Recognition Dataset- UCF50, which has 50 action categories and an average of 133 films per action category made up of authentic YouTube videos. This dataset is an expansion of the 11 action categories in the YouTube Action data collection (UCF11). Due to the variety of activity categories from its five primary types—playing musical instruments, sports, human-object contact, and bodily motion alone —UCF50 is a significantly more difficult dataset.

There are other performances in other categories, including sports, where most games are done with a greenery backdrop. Some of the films were shot from various angles, with various lighting conditions, and in various stances. The majority of the action recognition datasets on the market are staged and not realistic. The main goal of this dataset is to give the computer vision community access to a realistic action recognition dataset made out of YouTube videos.

Large changes in camera motion, object look and attitude, object scale, perspective, a crowded backdrop, lighting, and many other characteristics make this dataset exceedingly difficult to work with. Each of the 25 groups that make up the videos features more than five action sequences, for each of the 50 categories. Videos belonging to the same category could have elements in common, such as the same subject, a comparable setting, a comparable point of view, etc. As illustrated in figure 2 below, are some of the YouTube activity categories included in the UCF50 dataset: Drumming, Taichi, Bench Press, swing, Billiards Shot, Diving, Fencing, diving, etc.


Figure 2: Sample Action Categories of UCF50

## 3.4 DATA PREPROCESSING

Data pre-processing is used to prepare the data; all photos must be shrunk to a set size before being provided to CNN. We first remove the largest square we can from the image's centre, and we then enlarge each frame to the required resolution. To avoid accidentally distorting the dimensions of the persons in the video, we crop the video before resizing it. To prevent extra calculation, the prepared data is scaled with reference to a set width and height. After that, it is resized and normalised to provide data with a range of 0 to 1. Each variable is given an equal weight through normalisation so that no one variable may influence model performance only because they have larger numbers.

After then, a series of frames is created from the input videos. The specified sequences of frames will then be used to choose these sample frames. The acquired data is parsed during pre-processing into the input shape format for each model. Pre-processing improves the quality of the data to facilitate the discovery of valuable insights from the data. Additionally, it divides the dataset into 25% for testing and 75% for training.

## 3.5 DATA PREPARATION AND FEATURES EXTRACTION

For the representation and processing of pictures, CNN is the dominating source. When dealing with video data, CNN characteristics reflect each individual frame and then LSTM is used to determine the information that lies between each frame. These networks (CNN+RNN) are combined into a single model called LRCN for feature extraction and classification (Long Term Recurrent Convolutional Network). The next part talks about the thorough explanation. A video is made up of a series of frames that move between 20 and N times per second. The processing of the twenty to fifty redundant frames that make up a unit of time requires costly computations.

We will skip six frames while analysing a video for action detection because of this processing complexity. The trials clearly show that a six-frame hop has no impact on how the action is presented. Figure 3 illustrates the features representation situation. The first row in the figure depicts the frames in a series, and the second row displays the features maps for those frames. A little shift in the players' posture and orientation may be seen as one player tries to put the ball into the hoop.

CNN records all the minute changes in each frame because it can identify hidden patterns in images. To learn these adjustments sequentially to recognise actions in videos, RNN is employed.

Figure 3: Representation of frame order shift from frame to frame.

It takes thousands of photos to train a deep learning model for image representation, and the CNN model's weight modification requires a lot of computing power, such as a GPU. As a result, extraction of features and there categorizarion are done using the LRCN model.

The LRCN model consists of dropout layers in between each of the four convolution layers, four Maxpooling layers, and one fully connected layer. ReLU nonlinear activation function is added after each layer, and then an LSTM layer for classification and a softmax layer for action prediction are added at the last steps. The FC layer's retrieved features vector has 1,000 dimensions. Each frame's retrieved characteristics are regarded as one chunk for the RNN's input phase. RNN is fed CNN pieces for time intervals. As a result, we process six frames out of thirty for a second with a jump of six frames in a video. The RNN processes the six frames' worth of features in six separate pieces.

## 3.6 LRCN MODEL

This project uses Keras and TensorFlow to create a Long-Term Recurrent Convolutional Network (LRCN) model. The LRCN model combines the CNN and LSTM layers into a single model. It has a recurrent sequence model with a feature extractor (CNN) that can learn to comprehend the sequential input (LSTM). To extract features for our LRCN model, a 2DCNN model is used, and the output is then sent to the LSTM module. There are four Convolution layers, four Maxpooling layers, followed by dropout layer which is to reduce overfitting of the model. The resultant 2-Dimensional arrays from the combined feature maps are then flattened to create a single, substantial continuous linear vector. The flattened matrix is sent into the dense layer, which then classifies the image. The LSTM layer will sequentially process the information for classification and at last SoftMax is used which will gives decimal probability to each class. The sum of their decimal probabilities must be 1.0. In the picture 4 below, the system architecture is depicted.

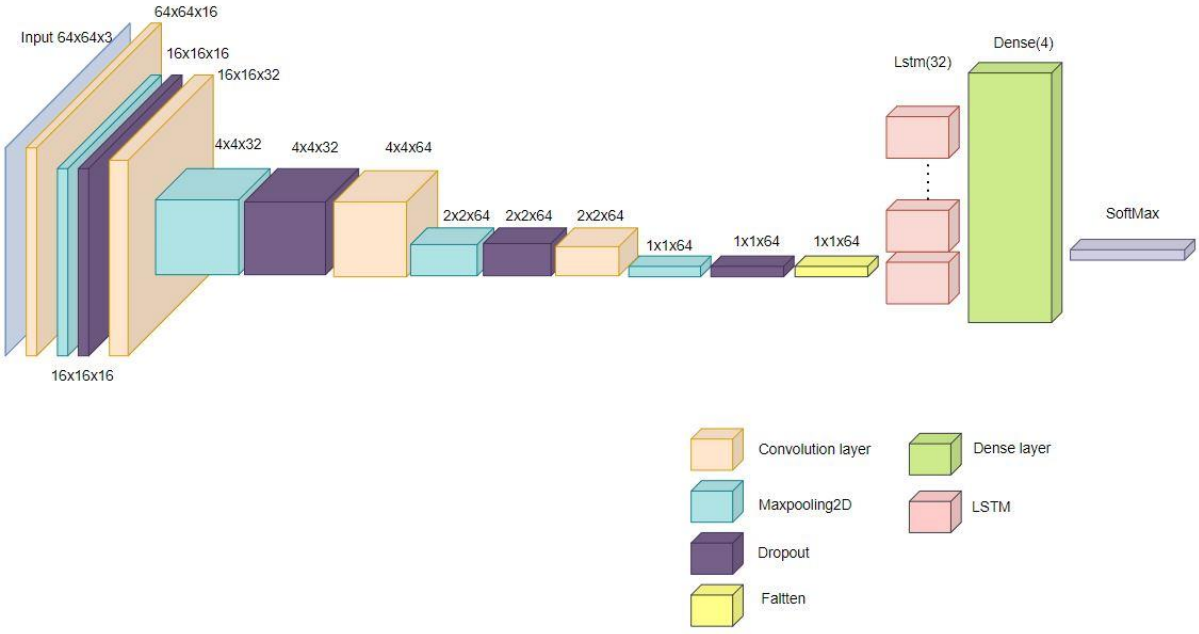


Figure 4. Architecture of LRCN model

The CNN model may be tuned for the job at hand to extract spatial information from the video's frame-by-frame data, either using a pre-trained model or a regular CNN. The LSTM model may then utilise the CNN properties to predict the behaviours shown in the video. The convolutional layers are used to provide the spatial data from the frames to the LSTM layer(s) at each time step in order to model the temporal sequence.

The network directly learns how to cope with spatiotemporal input during end-to-end training, which results in a reliable model. The final flattening of the model's output layers feeds the SoftMax-activated Dense layer, which outputs the probabilities for each action category. Dropout layers are further employed to avoid the model fitting the data too closely and MaxPooling3D layers to reduce frame size and get rid of extraneous calculations.

It is possible to better comprehend the context of an event when watching video since visual motions are represented by a number of frames. RNNs could be able to comprehend lengthy sequences, but they would likely forget the sequence's starting inputs. The vanishing gradient issue can be resolved with an RNN variation called LSTM (Long Short-Term Memory).

It can pick up enduring dependencies. Input, output, and forget gates make up its structure, which controls how long-term sequence patterns are recognised. A sigmoid unit that adjusts the gates learns how to open and shut them during training adjusts the gates. The action label will be generated by the model for each type of activity. The LRCN model is seen in Figure 5 below.

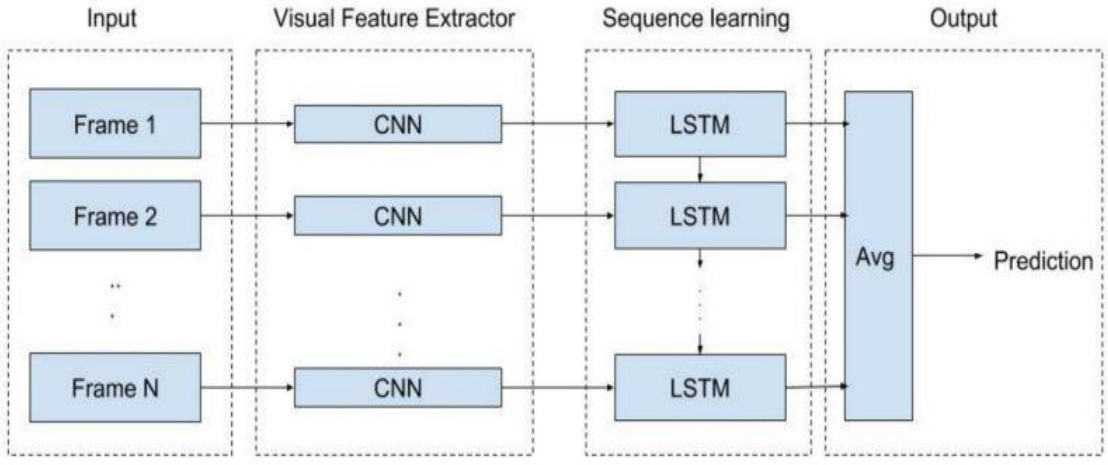


Figure 5: LRCN model

Additionally, TimeDistributed wrapper layers are employed, allowing for the independent application of the same layer to each individual temporal frame of the video. Each input must have a minimum of three dimensions, with the temporal dimension being the index-one dimension of the first input. Due to this, if the layer's original input shape was (width, height, number of channels), it enables the layer (around which it is wrapped) to accept input of the shape (no of frames, width, height, number of channels), which is very beneficial since it enables the single-shot entry of the complete video into the model. The figure 6 below displays the TimeDistributed layer.

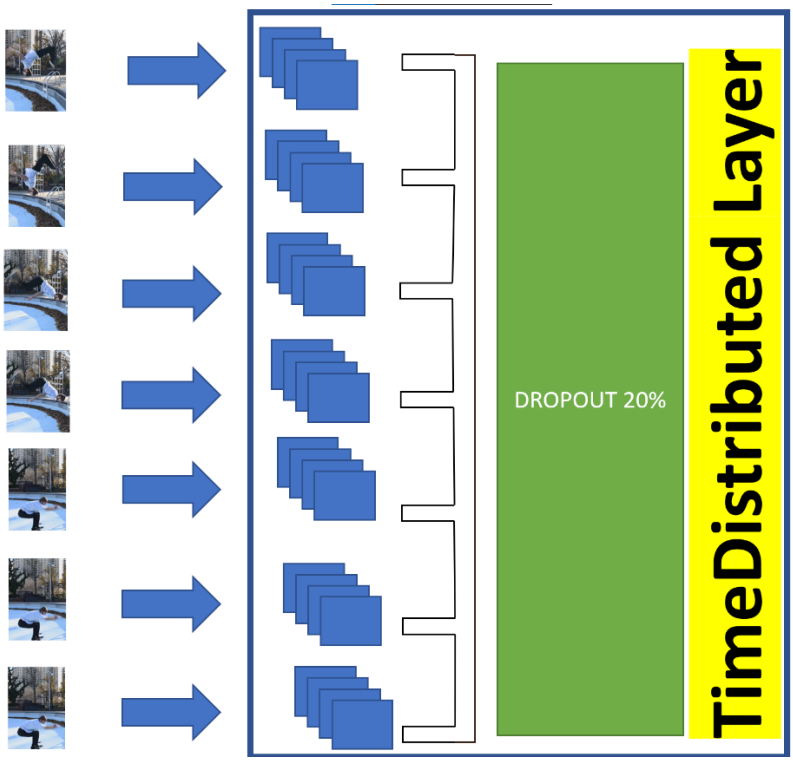


Figure 6: Time distributed layer

The overall structure of the model created is shown as figure 7 below.

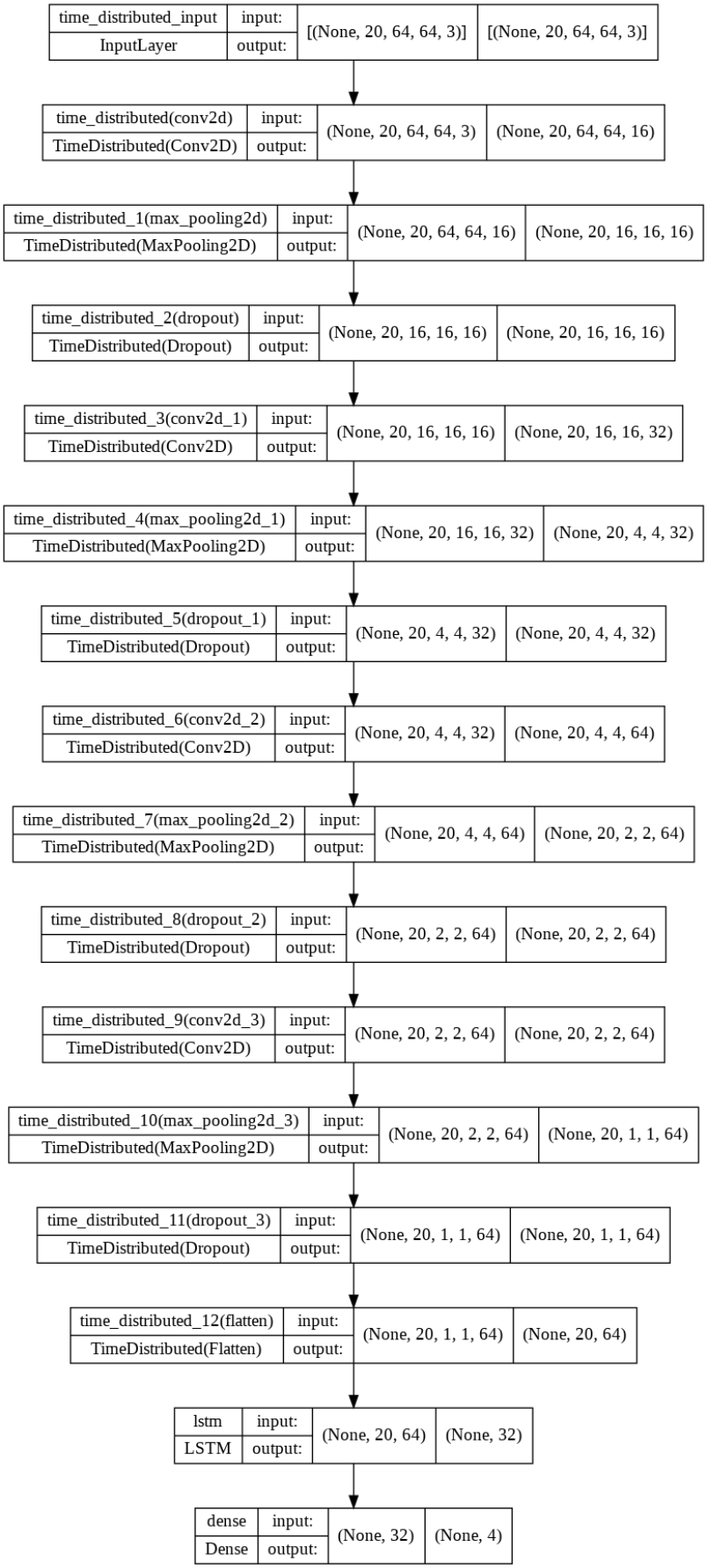


Figure 7: Model structure

## 3.7 ALOGRITHMS USED

### 3.7.1 CNN

An artificial neural network called a convolution neural network (CNN) is employes for image processing and recognition. CNNs are created primarily to examine the pixel input. Among the applications of machine vision are recommender systems, natural language processing, and picture and video recognition, and image processing are among them. Deep learning is also often used in artificial intelligence to do generative and descriptive tasks (NLP). A CNN's layer structure consists of three levels: an input layer, an output layer, and a hidden layer, as well as many convolutional layers, pooling layers, fully connected layers, and normalising layers.

**2DCNN**

Based on the size of the convolutional kernel that are used, CNNs for segmentation may be classed. 2D convolutional kernels or filters are utilised by 2D CNN to forecast the segmentation map for a particular slice. By considering predictions one slice at a time, segmentation maps are predicted as a whole. The context throughout the height and breadth of the slice may be used by the 2D convolutional filters to produce predictions. To extract characteristics from frames during treatment, conventional Convolutional Neural Network (CNN) and Max Pooling approaches are often used over all frames. The deeper the network, the more computationally expensive the model is, even if CNN layers are capable of detecting high level characteristics from input frames across a stack of layers.

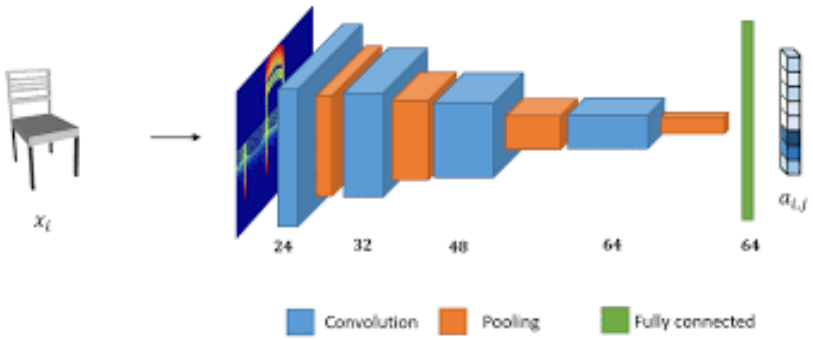


Figure 8: Architecture of 2DCNN

### 3.7.2 LSTM

Recurrent neural network (RNN) models called LSTM network models have the capacity to learn and remember across lengthy input data sequences. They are designed to be used with data that comprises of several time-steps, up to 200 or 400. RNNs are made up of linked neurons with input, internal (or hidden), and output units that are activated at time t and may process data sequentially and selectively. It can simulate outputs made up of sequences of non-independent elements since it processes one element at a time.

The RNN design supports time-space data such as audio, video, and text analysis and the discovery of hidden patterns. Biases and weights are applied to the data before it is supplied to activation functions. Because of the numerous calculations, the original inputs' influence on the subsequent series of data becomes insignificant after a few layers, which causes the vanishing gradient issue. LSTM is the answer to this issue. The memory cell, input gate, output gate, forget gate, and non-linear gating devices that monitor the information flow into and out of the cell are the primary components of the LSTM architecture, which can preserve its state over time.

Because LSTMs may directly learn from the unprocessed time series data, they are advantageous for sequence categorization. The model should perform similarly to models fitted on a version of the dataset with engineered features and be able to internalise the time series data. It can analyse long videos by looking at particular aspects for a certain amount of time.

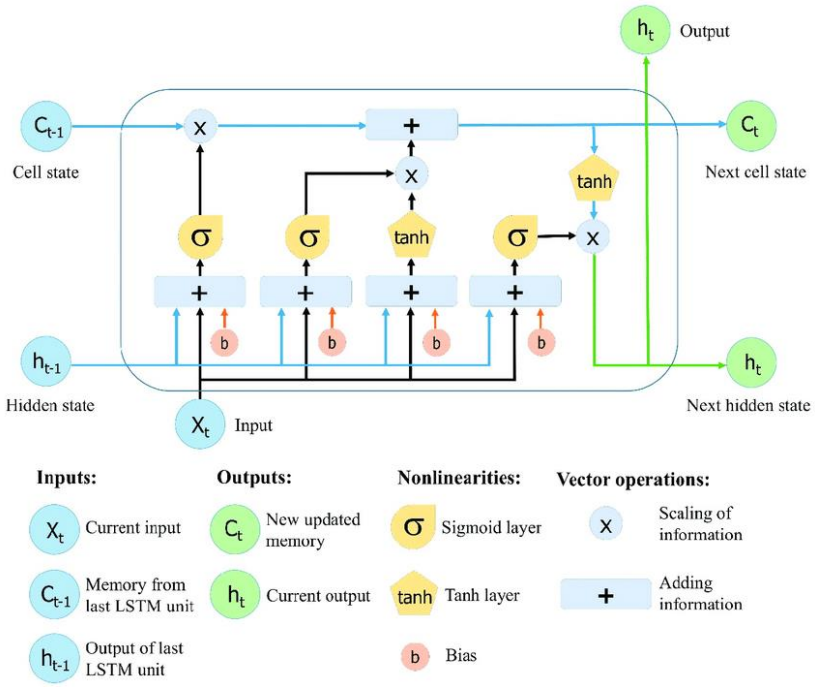


Figure 9: Architecture of LSTM

# Chapter 4

# Result and Discussion

## 4.1 Performance and Evaluation

The experimental results are discussed after utilising the UCF50 dataset as a common benchmarking dataset based on the accuracy requirement then an explanation of the results of the experiment follows. The dataset is sectioned into training and validation sections, with the respective proportions of 75% and 25%. Python 3 and TensorFlow, two deep learning frameworks, are utilized to implement the suggested model. The combination of naturally occurring realistic activities and the behaviours portrayed by several actors in the dataset UCF50 is one of the dataset's biggest problems.

A total of 487 videos are trained from the dataset. This recognition system has achieved an accuracy of 95% on UCF50 dataset. The validation loss and validation accuracy are labelled below table.

| Accuracy | Validation_loss | Validation_accuracy |
|---|---|---|
| 0.9555 | 0.3219 | 0.9041 |

Table 1: Accuracy on the model

The CNN model built on LRCN learns more quickly than previous models, reaching a maximum validation accuracy of 90% after 43 epochs and a training accuracy of about 95% after 40 epochs. Additionally, this model does not exhibit over-fitting, as can be shown. 100 epochs are used to train the model, using a dropout of 0.25 and a four-batch size. Table 2 as shown below includes a list of the training parameters.

| | |
|---|---|
| Total Parameters Learned | 73,060 |
| Types of custom layers used | Convolutional, Maxpooling2d, dropout, flatten, dense, SoftMax |
| Input image size | 64x64 |
| Activation function | Relu |
| Optimizer | Adam |
| Error estimate | categorical_crossentropy |
| Batch size | 4 |
| Training epochs | 100 |

Table 2: Training parameters

The LRCN model's training-validation accuracy and training-validation loss are shown in Figures 10 and 11, respectively. It is obvious that these models were successful in generalisation and absorbed the characteristics of the training set.

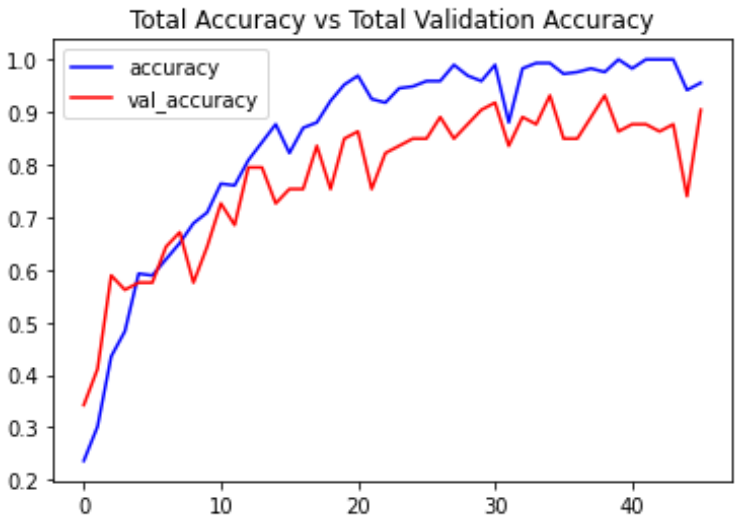


Figure 10: Accuracy curve for Training-validation

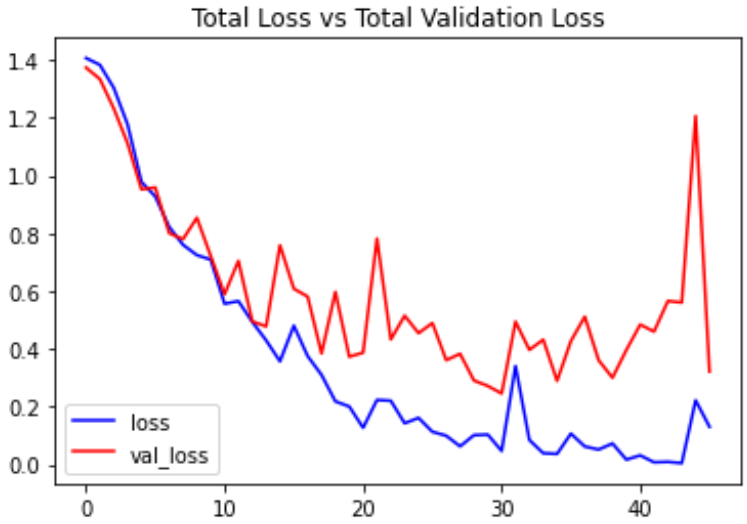


Figure 11: Loss Curve for Training-validation

# Chapter 5

# Conclusion

In this study, an approach for predicting human actions in videos has been put forth. In order to achieve these goals, the LRCN model has been employed in order to extract features and categorize them. The suggested technique learns long-term complicated sequences in videos by first extracting CNN characteristics from the video frames and feeding them to LSTM. Next, the SoftMax function was used to label and categorize the actions, and both of these networks are combined into a single model. As a result, complex frame-to-frame sequential patterns in the features were simpler to spot.

The frequently used UCF dataset was utilised to assess the performance of the systems. The URL of a test video from YouTube was provided to gauge how well the model recognised the video. The accuracy metrics were examined, and the overall findings demonstrated that the suggested strategy could generalise effectively. The system has achieved an accuracy of 95%. I took considerably less time for the whole recognition purpose. By simplifying the computing process, this system was able to attain great accuracy. Potential academics are already interested in action recognition due to its many applications, such in smart homes, security systems, content filtering, etc. It even allows for remote observation of individuals at work, in their homes, or in public areas.

## 5.1 FUTURE ENHANCEMENT

This might be developed further in the future to enhance the architecture and forecast future actions based on current actions as well as semantic scene segmentation and comprehension. Several benchmark action recognition datasets may be used to contain more activities. The detailed details of each action might likewise be better captured using certain complicated networks, in a similar manner. Classification, the creation of video captions, and summarising might all be accomplished using this action recognition.

# References

[1] "Human Action Recognition Using Deep Learning Methods on Limited Sensory Data," IEEE Sensors Journal, vol. 20, no. 6, March 15, 2020, pp. 3101-3112, doi: 10.1109/JSEN.2019.2956901. N. Tufek, M. Yalcin, M. Altintas, F. Kalaoglu, Y. Li, and S. K. Bahadir.

[2] "HAR-Depth: A Novel Framework for Human Action Recognition Using Sequential Learning and Depth Estimated History Images," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 5, no. 5, Oct. 2021, pp. 813–825, doi: 10.1109/TETCI.2020.3014367 by S. P. Sahoo, S. Ari, K. Mahapatra and S. P. Mohanty.

[3] "Movie tags prediction and segmentation using deep learning," in U. A. Khan, M. A. Martinez-Del-Amor, S. M. Altowaijri, A. Ahmed, A. U. Rahman, N. U. Sama, K. Haseeb, and N. Islam, IEEE Access, vol. 8, pp. 6071–6086, 2020.

[4] "Asymmetric 3D convolutional neural networks for action detection," Pattern Recognit., vol. 85, pp. 1–12, Jan. 2019, by H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank.

[5] "ImageNet classification using deep convolutional neural networks," A. Krizhevsky, I. Sutskever, and G. Commun. ACM, 60, no. 2, June 2012, pp. 84–90.

[6] Large-scale video categorization with convolutional neural networks was described by A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., June 2014, pp. 1725–1732.

[7] Learning spatiotemporal features with 3D convolutional networks, Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4489-4497. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri.

[8] Asymmetric 3D convolutional neural networks for action recognition, H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, Pattern Recognit., vol. 85, pp. 1–12, Jan. 2019.

[9] "Human action detection using transfer learning with deep representations," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), May 2017, pp. 463-469. A. B. Sargano, X. Wang, P. Angelov, and Z. Habib.

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

[11] A combined multiple action recognition and summarization for surveillance video sequences, International Journal of Speech Technology, vol. 51, no. 2, pp. 690–712, February 2021. O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi.

# Appendices


Figure 12: Output1 of system


Figure 13: Output2 of system