

Human Action Recognition Using Deep Learning

Satyam Kumar¹, Shubham Kumar², Laxmi V³

¹Student, Dept. of Computer Science and Engineering, BNM Institute of Technology, Karnataka, India

² Student, Dept. of Computer Science and Engineering, BNM Institute of Technology, Karnataka, India

³ Assistant Professor, Dept. of Information Science and Engineering, BNM Institute of Technology, Karnataka, India

Abstract – The goals of video analysis tasks have changed significantly over time, shifting from inferring the current state to forecasting the future state. Recent advancements in the fields of computer vision and machine learning have made it possible. Different human activities are inferred in tasks based on vision-based action recognition based on the full motions of those acts. By extrapolating from that person's current actions, it also aids in the prognosis of that person's future action. Since it directly addresses issues in the real world, such as visual surveillance, autonomous cars, entertainment, etc., it has been a prominent topic in recent years. To create an effective human action recognizer, a lot of study has been done in this area. Additionally, it is anticipated that more work will need to be done. In this sense, human action recognition has a wide range of uses, including patient monitoring, video surveillance, and many more. Two CNN and LRCN models are put out in this article. The findings show that the recommended approach performs at least 8% more accurately than the traditional two-stream CNN method. The recommended method also offers better temporal and spatial stream identification accuracy.

Key Words: Convolutional Neural Network (CNN), Long-term Recurrent Convolutional Network (LRCN), Human Activity Recognition (HAR), Deep Learning (DL).

1. INTRODUCTION

Human-Human interactions, such as embracing and shaking hands, as well as Human-Object interactions, like playing a guitar or tossing a ball, are the two categories in which human action can be broadly classified. This categorization is based on human behavior, and human behavior is defined by gestures, positions, etc. Due to differences in speed, lighting, partial occlusion of persons, perspective, and anthropometry of those taking part in the different interactions, it might be difficult to recognize human activity. Others are more strongly connected to the process of recognizing the activity, while certain identification issues are more closely related to the difficulty of identifying and monitoring people in videos. Action classes are vulnerable to substantial intra-class variance because of the camera's perspective. When a person raises an arm while being observed from the front or the side, it is not the same gesture as when the same arm is raised. Additionally, lifting the left, right, or both arms in different ways all denote the same motion of

raising the hands. Then there is the issue of a person's particular way of performing a gesture, including how long it lasts and how it is performed. Another thing to consider is human anthropometric variations, such as those brought on by age, size, and body component ratios.

2. RELATED WORK

Over the last few decades, researchers have presented a number of handmade and deep-nets-based action detection algorithms. In the early effort, which was based on hand-crafted features for unrealistic activities, an actor would carry out various actions in a scenario with a simple background. These systems extract low-level characteristics from the video data and transmit them to a classifier like a support vector machine (SVM), decision tree, or KNN in order to detect activities.

Abdellaoui, M., Douik, A. et al,[1] proposed a brand-new DBN-based HAR technique, attempts to increase the accuracy of human activity categorization. Segment the video clips from the human activity dataset into frames as a first step. The result is then converted to binary frames, and the resulting frames are subjected to a series of morphological filtering techniques in order to improve their quality. Then, generate an input matrix containing the training data, the testing data, and their labels by converting the new frames into a binary vector. The input data for our DBN architecture are represented by this matrix. The DBN classifier will be trained using the training data matrix in the last phase, and the classification outcome will be extracted.

Yu-Wei Chao et al,[2] proposed a better method for locating temporal activity in video, called TAL-Net, and was modelled around the Faster RCNN object identification framework. By using a multi-scale architecture that can accommodate extreme variation in action durations, TAL-Net can improve receptive field alignment, extend receptive fields to more effectively exploit the temporal context of actions for proposal generation and action classification, and explicitly take into account multi-stream feature fusion while highlighting the significance of fusing motion late. On the THUMOS'14 detection benchmark, the model achieves state-of-the-art performance for both action suggestion and localization, as well as competitive performance on the Activity Net challenge.

For video recognition, Feichtenhofer, Christoph & Fan, et al,[3] introduce SlowFast networks. The suggested model includes two pathways: (i) a Slow pathway that operates at a low frame rate to record spatial semantics, and (ii) a Fast pathway that operates at a high frame rate to record motion with precise temporal precision. It is feasible to make the Fast route exceedingly light while still enabling it to acquire important temporal information for video identification by reducing the channel capacity. The SlowFast technique is acknowledged with generating considerable improvements, as the proposed models successfully classify and detect actions in films. On the crucial video recognition tests Kinetics, Charades, and AVA, provide cutting-edge accuracy.

The automatic detection of human behaviors in surveillance footage is explored by Ding, Chunhui & Fan, et al,[4]. The majority of currently used methods rely their classifiers on very complex characteristics that are computed from the raw inputs. CNN is a deep model that can respond immediately to the raw inputs. However, as this time, these models can only accept 2D inputs. In this study, a brand-new 3D CNN model for action recognition is developed. This model utilizes 3D convolutions to extract features from the spatial and temporal dimensions, effectively capturing the motion information recorded in a large number of nearby frames. The developed model generates multiple channels of information from the input frames, and the final feature representation combines information from channels. further, boost the performance and propose regularizing the outputs with high-level features and combining the predictions of a variety of different models.

Convolutional Neural Networks (CNNs), a strong group of models for image recognition issues, have been proposed by Andrej Karpathy et al,[5]. Encouraged by these findings, present a thorough empirical analysis of CNNs' performance on large-scale video classification using a new dataset consisting of 1 million YouTube videos divided into 487 classes. Investigate several ways to strengthen a CNN's connection in the time domain so that it can take use of local spatiotemporal data, and suggest a multiresolution, foveated architecture as a potential fast-tracking option. By retraining the top layers on the Kinetics-400 Action Recognition dataset and seeing notable performance increases compared to the Kinetics-400 baseline model, further research the generalization performance of our best model (63.3 percent up from 43.9 percent).

3. MOTIVATION

A variety of applications, including video surveillance, identification verification, the development of intelligent systems for human-machine interaction, and many more, benefit from the usage of human action recognition (HAR). In the realm of computer vision and machine learning, it is

a difficult problem. Recognizing human action requires an effective feature representation. For HAR, identifying characteristics and focusing only on the geographical region is insufficient; it also has to consider how features have changed over time. Recent years have seen the development of a wide range of action representation techniques, including local and global features based on temporal and spatial changes, trajectory features based on key point tracking, motion changes based on depth information, and action features based on changes in human pose. Due to its effectiveness in object identification and picture categorization, deep learning has been widely employed by researchers to identify human action.

4. PROPOSED METHOD

The proposed methodology considers two different deep learning models to predict action in the video. After applying the models for prediction, determination of which model performs better and select an appropriate model to obtain better efficacy. In this paper, the proposed models are CNN and LRCN, which are two of the most recent and effective methods for action prediction in videos, as deep learning mechanisms. The process design of the suggested approach is shown in Figure 1.

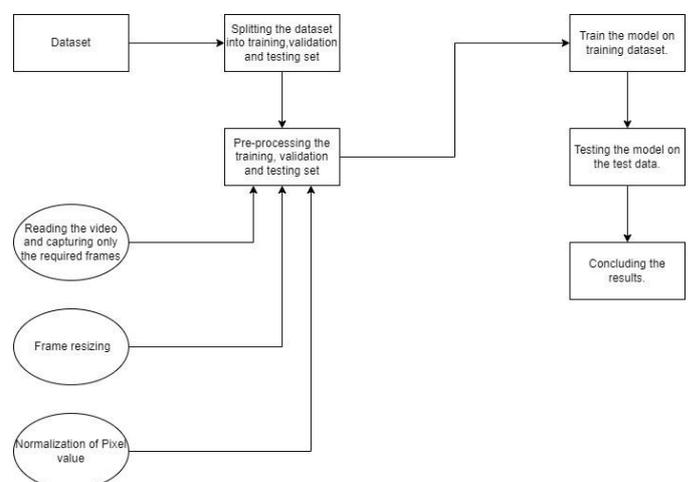


Fig -1: Illustration of proposed method

4.1 Dataset and Dataset Preprocessing

Dataset: The Kinetics dataset is a sizable, top-notch dataset for identifying human activity in video. Depending on the dataset version, the Kinetics 400[1] collection of large-scale, high-quality datasets of URL links to up to 650,000 video clips covers 400/600/700 human activity classes. Handshakes, embraces, and other human-human and human-object interactions are shown in the films, along with musicians playing instruments. Each action class has at least 400/600/700 videos. The average length of each clip is 10 seconds, and each has been manually assigned a single action type.

Dataset Preprocessing: The Kinetic-400 dataset will be imported first then, dataset has been divided into three sets: training set, test set, and validation set in ratios of 0.8, 0.1, and 0.1 respectively. In a pre-processing step, first, extract a frame from video, then the frame will be resized into 64 x 64 and at last normalization of pixel value occurs. After preprocessing of the dataset, the model CNN and LRCN will be created and the training set will be fed into the model, after that the model will test the model on test data. At last, the result will be concluded which model is better by comparing accuracy.

4.2 CNN

A kind of deep neural networks called convolutional neural networks (CNN/ConvNet) are most frequently used to assess visual images. When people think about neural networks, matrix multiplications typically spring to mind, but with ConvNet, that is not the case. It makes use of a unique method called convolution. In mathematics, convolution shows how the form of one function is altered by the other by taking two functions and producing a third function. The picture is divided into regions, and each area is then assigned a different hidden node. Only one area of the image is where a pattern is discovered by each concealed node. This region is under the authority of a kernel, often known as a filter or window. A kernel technique allows us to apply linear classifiers to non-linear problems by projecting non-linear data into a higher-dimensional space without physically visiting or comprehending that higher-dimensional region.

Convolved across the x- and y-axes is a filter. Several filters are used to extract different patterns from the image. A variety of filters are utilized at various tiers when building the model. In our model, 4,8,16 different filters are applied at various levels. When the output of one filter is convolved across the whole picture, a 2-d layer of neurons known as a feature map is created. Each filter manages a single feature map. Each filter manages a single feature map. These feature maps may be layered to create a three-dimensional array, which may then be utilized as an input to other layers. The Convolutional layer of a CNN is responsible for doing this. Following these layers are the pooling layers, as depicted in Figure 2, which minimize the output's spatial dimensions (obtained from the convolutional layers). Simply said, a window is shifted in both directions, and the filter or window's maximum value is applied (Max-Pooling layer). The sole distinction between the two is that, on occasion, the average value during the timeframe is utilized in place of the highest number. As a consequence, the convolutional layers increase the depth of the input picture while the pooling layers diminish the spatial dimensions (height and width). A image that can be flattened into a 1-dimensional array has information that is encoded using a significant architecture.

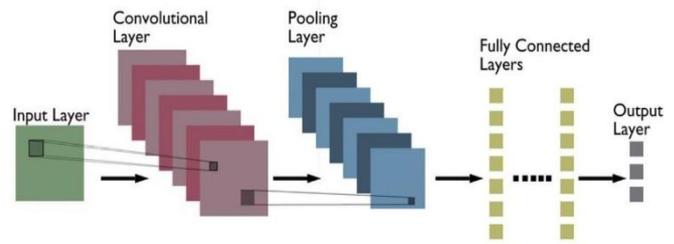


Fig -2: CNN Architecture

4.3 LRCN

The end-to-end trainable LRCN is a CNN and LSTM combination that is appropriate for complex visual comprehension tasks including activity detection, video description, and picture captioning. When compared to earlier models, which either assumed a fixed visual representation or carried out simple temporal averaging for sequential processing, recurrent convolutional models are deeper because they learn compositional representations in space and time. Additionally, because it is directly coupled to the convolutional network, it may be taught in temporal dynamics and convolutional perceptual representations.

As seen in figure 3, LRCN uses a CNN to handle the variable-length visual input. Additionally, the LSTM, a stack of recurrent sequence models, receives input from their outputs. The final result of the sequence models is a variable-length prediction. With time-varying inputs and outputs, such as activity detection, picture captioning, and video description, the LRCN is a suitable model for handling these tasks.

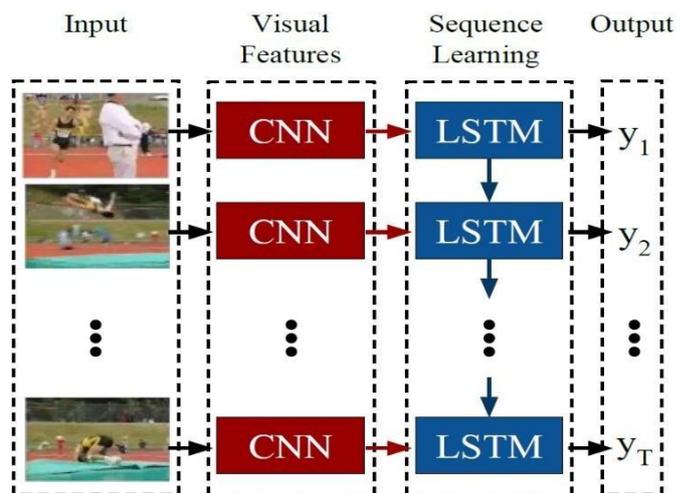


Fig -3: LRCN Architecture

5. RESULTS AND DISCUSSION

Results from CNN and LRCN Models: The implementation was completed using the Python

programming language in the Google Collaboratory's Jupyter notebook environment. Model when run on video, is able to recognize Human action well. The frames are extracted from video. It is processed and passed through the model. Then the model predicts the human action. Generalization hence can be taken as another metrics for validation and evaluation of the performance. It is not something that can be plotted, but the evaluation can be done in terms of human action recognition. The several factors affecting this can be the background clutter, changes in scale, lighting and appearance and so on. As discussed, the models are trained with more than 300 videos classes. Thus, it can easily predict human action accurately. The total loss together with validation loss and training accuracy along with validation accuracy varied over iterations of performance evaluation for the two models as shown in below figure 4 - 7. In epoch versus accuracy graph, epoch is plotted on x-axis and accuracy is plotted on y-axis, when epoch is increased the accuracy also gets increased, while in epoch versus loss graph, epoch is plotted on x-axis and loss is plotted on y-axis, when epoch is increased the loss gets decreased.

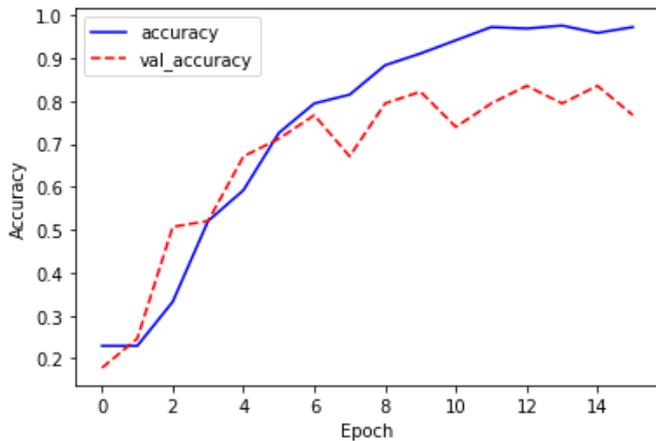


Fig -4: With the CNN model, overall and validation accuracy.

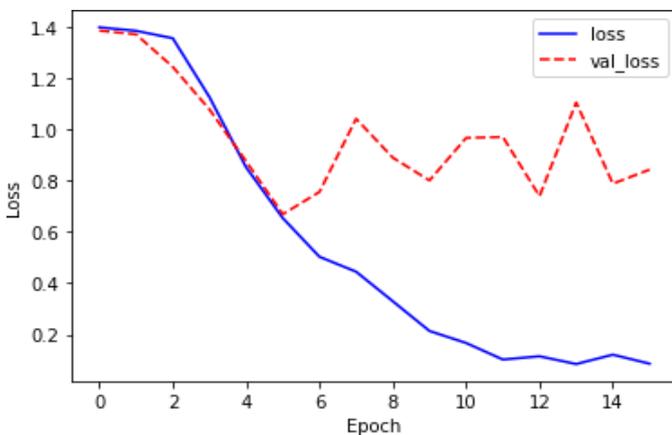


Fig -5: With the CNN model, overall and validation loss.

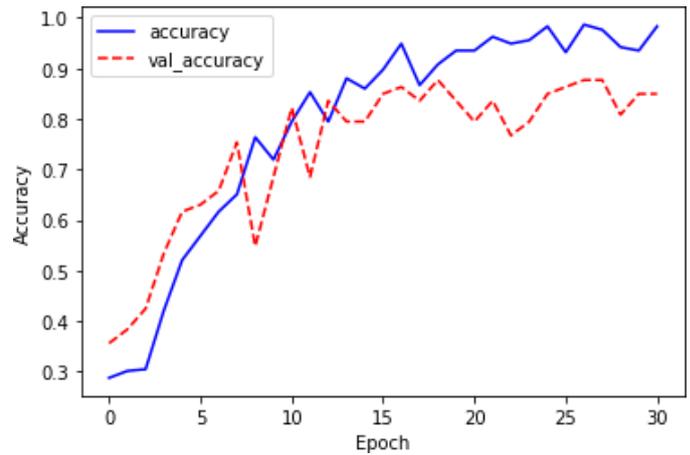


Fig-6: With the LRCN model, overall and validation accuracy.

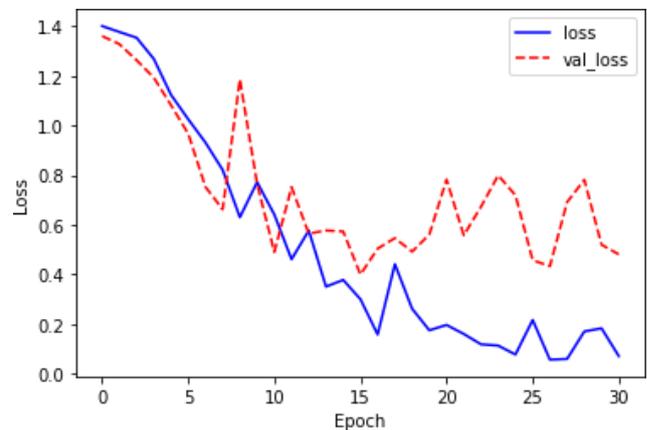


Fig -7: With the LRCN model, overall and validation loss.

Table 1, shows the results of action recognition in different videos. When video is provided as input, the frame will be extracted from the video and fed into the model. Model extract different features from frame and label action to frame and labeled frame will be displayed as output.

Table 1: Different video Human action Recognition Results.

File Name	Input Image	Result
Playing_Piano.mp4		This video is provided as an input and model identified the action as playing piano.
Archery.mp4		This video is provided as an input and model identified the action as Archery.

Making_pizza.mp4		This video is provided as an input and model identified the action as Making Pizza.
Performing_yoga.mp4		This video is provided as an input and model identified the action as Performing Yoga.
Playing_flute.mp4		This video is provided as an input and model identified the action as Playing Flute.
Javelin_throw.mp4		This video is provided as an input and model identified the action as Archery.
Man_running_on_treadmill.mp4		This video is provided as an input and model identified the action as Man running on treadmill.
Person_stretching_leg.mp4		This video is provided as an input and model identified the action as Person stretching leg.
Child_climbing_on_tree.mp4		This video is provided as an input and model identified the action as Child climbing on tree.
Person_jogging_on_street.mp4		This video is provided as an input and model identified the action as Person Jogging.
Air_drumming.mp4		This video is provided as an input and model identified the action as Air

		Drumming.
Abseiling.mp4		This video is provided as an input and model identified the action abseiling.
Person_performing_triple_jump.mp4		This video is provided as an input and model identified the action as Person Performing triple jumping.
Person_making_tea.mp4		This video is provided as an input and model identified the action as Person making tea.

Comparison of CNN and LRCN Model Results: Create a classification report first, which includes the results obtained for the two models. After comparing the results obtained in paper Action Recognition in Video Sequences using Deep Bi-directional LSTM with CNN Features [12], our model outperforms by at least 8% accuracy. Finally, it has been determined that the relative to the CNN model, the LRCN model displays much higher accuracy. Comparison of Results from CNN and LRCN Models shown in Table 2.

Table 2: Comparison of Results from CNN and LRCN Models

Model	CNN Model	LRCN Model
Accuracy	84.68%	89.71%
Average Precision	80%	85%
Average Recall	79%	86%
Average F1-Score	80%	83%

6. CONCLUSION AND FUTURE SCOPE

A CNN model and an LRCN model are described in this study. The suggested technique seeks to extract both spatial and temporal characteristics from the RGB video frames in the spatial stream, in contrast to the conventional two-stream CNN model. This goal is

accomplished by using an LSTM-based model in place of the conventional convolutional neural network in its spatial stream. The CNN model has been made fast and robust in terms of speed and accuracy with the use of Conv2D layers, dropout regularization, and ideal model hyperparameters. The model proposed in this paper obtained, 84.68% accuracy and 89.71% accuracy in CNN and LRCN model respectively with the Kinetic-400 dataset. In this article adopting this Human Activity Recognition framework as a future effort to develop an IoT-based smart monitoring system for eldercare or childcare. Additionally, if one can create our own dataset using a camera for a certain set of common activities individuals engage in on a daily basis, that would be ideal for the task. Deep Learning applications seem to have a wide range of enhanced uses in the near future for this study field. The use of the reinforcement learning paradigm on the area of activity recognition and classification is also suggested in the paper as a potential direction for future development.

REFERENCES

- [1] Abdellaoui, M., Douik, A. (2020). Human action recognition in video sequences using deep belief networks. *Traitement du Signal*, Vol. 37, No. 1, pp. 37-44. <https://doi.org/10.18280/ts.370105>
- [2] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, Rahul Sukthankar; Rethinking the Faster R-CNN Architecture for Temporal Action Localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1130-1139
- [3] Feichtenhofer, Christoph & Fan, Haoqi & Malik, Jitendra & He, Kaiming. (2018). SlowFast Networks for Video Recognition.
- [4] Ding, Chunhui & Fan, Shouke & Zhu, Ming & Weiguo, Feng & Jia, Baozhi. (2014). Violence Detection in Video by Using 3D Convolutional Neural Networks. 8888. 551-558. [10.1007/978-3-319-14364-4_53](https://doi.org/10.1007/978-3-319-14364-4_53).
- [5] Karpathy, Andrej & Toderici, George & Shetty, Sanketh & Leung, Thomas & Sukthankar, Rahul & Fei-Fei, Li. (2014). Large-Scale Video Classification with Convolutional Neural Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1725-1732. [10.1109/CVPR.2014.223](https://doi.org/10.1109/CVPR.2014.223).
- [6] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16– 28, Jul. 2019.
- [7] Y. Kim and H. Ling, "Human activity classification based on microDoppler signatures using an artificial neural network," in *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, Jul. 2008.
- [8] Jalal A, Kamal S, Kim D. A Depth Video Sensor-Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor Environments. *Sensors*. 2014;14(7):11735-11759. <https://doi.org/10.3390/s140711735>.
- [9] J. Li, S. L. Phung, F. H. C. Tivive, and A. Bouzerdoum, "Automatic classification of human motions using Doppler radar," in *Proc. Int. Joint Conf. Neural Netw.(IJCNN)*, Jun. 2012, pp.1– 6.
- [10] MW. Li, B. Xiong, and G. Kuang, "Target classification and recognition based on micro- Doppler radar signatures," in *Proc. Progr. Electromagn. Res. Symp.-FALL (PIERS-FALL)*, Nov. 2017, pp. 1679–1684
- [11] <https://www.deepmind.com/open-source/kinetics>
- [12] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," in *IEEE Access*, vol. 6, pp. 1155-1166, 2018, doi: [10.1109/ACCESS.2017.2778011](https://doi.org/10.1109/ACCESS.2017.2778011).