HUMAN ACTIVITY RECOGNITION USING CNN

A.MURUGESHWARI
PG Student
Master of Engineering-Embedded System Technologies
Department of Electronics and Communication Engineering
National Engineering College, Kovilpatti, India
Email-mrugeshwari1996@gmail.com

S. TAMIL SELVI
Professor
Department of Electronics and Communication Engineering
National Engineering College, Kovilpatti, India
Email-stsece@nec.edu.in

Abstract- Human activity acknowledgment (HAR) explore is hot in computer vision, however high exactness acknowledgment of human activity in the unpredictable foundation is as yet an open inquiry. In this paper, the deep model convolutional neural network (CNN) that can classifies HAR facilities directly on the raw inputs. To enable improved real-world applications an efficient pretraining strategy has been introduced to reduce the high computational cost of kernel training.

Keywords – Convolutional Neural Networks, deep learning, pre-training, Human Activity Recognition

I.INTRODUCTION

Human Activities can be classified in various categories including Human-Human interaction and Human-Object interaction. Human activtity recognition (HAR) is an important area of computer vision research and Applications. HAR classification is based on human activities,where activities of people are determined by their motions, presents and so on. The objective of the action recognition is a mechanized investigation or translation of continuous occasions and their setting from video data. The levels of human activities are single actor Movements, Interactions, group activities.HAR applications include patient monitoring systems, surveillance systems, and a variety of systems that involve inter actions between persons and electronic devices such as human-computer interfaces. The composed of multiple simple actions of persons , these applications require recognition of high-level activities. Here, an automatic Human activity recognition classification techniques using the idea of convolution neural network (CNN) is proposed.

II.RELATED WORKS

In paper [1] the author used a action bank features of UCF50 database to develop a deep convolutional network architecture for recognizing human actions in videos.

In paper [2] the author established 3D CNN models for action recognition .These models develop features from both spatial and temporal measurements by performing 3D convolutions. To perform convolution and subsampling seperatelty in each channel it has to developed deep architecture that generates multiple channels of information from adjacent input frames. The final feature portrayal is obtained by consolidating data from all channels.

In paper [3] the author focused on Human action recognition is a challenging task as the articulated action data is high dimensional in both spatial and temporal domains. A successful way to deal with handle this unpredictability is to isolate human body into various body parts as indicated by human skeletal joint positions, and performs acknowledgment dependent on these part-based element descriptors.

In paper [4] the author reviews reviews different strategies applied to transcribed character recognition and looks at them on a standard written by hand advanced recognition task Convolutional Neural Networks that are explicitly intended to manage the fluctuation of 2D shapes are to beat every single other strategy.

In paper [5] the author focused on sums up the ongoing advancement of activity acknowledgment from the start. At that point dependent on Hierarchical Filtered Motion model and Nearest Neighbor classifier, do activity acknowledgment utilizing Histogram of Oriented Gradients (HOG)include in video groupings of various goals. Here they use KTH dataset for preparing and MSR activity dataset s for testing. The examination shows that the new component extraction process is viable and has better execution in the cross-dataset activity acknowledgment.

In paper [6] the author presents a video portrayal dependent on thick directions and movement limit descriptors. Directions catch the nearby movement data of the video. A thick portrayal ensures a decent inclusion of frontal area movement just as of the encompassing setting. A best in class optical stream calculation empowers a hearty and proficient extraction of thick directions.

In paper [7] the author introduced, a constant following based way to deal with human activity acknowledgment is proposed. The technique gets as info profundity map information streams from a solitary kinect sensor. At first, a skeleton-following calculation is applied. At that point, another activity portrayal is presented, which depends on the estimation of circular points between joints and the particular precise speeds.

In paper [8] the author investigated they conjecture that the characterization of activities can be supported by planning a keen element pooling procedure under the commonly utilized pack of-words-put together representation. Founded with respect to programmed video saliency investigation, they propose the spatial-worldly consideration mindful pooling plan for highlight pooling.

CNN can learn suitable highlights by them naturally, which prompts great item acknowledgment and grouping precision. The rest of the papers are organized as follows: Section III describes the proposed method, Section IV includes the dataset preparation, Section V includes the implementation steps and Section VI describes the result and discussion. Scope for further research in this work is detailed in the final section.

III.PROPOSED METHOD

In this work, a multilayered CNN and Long Short-Term Memory model is used. To support sequence prediction the CNN LSTM architecture involves using Convolutional Neural Network (CNN) layers for feature extraction on input file combined with LSTMs .CNN LSTMs were developed for visual time series prediction problems and therefore the application of generating textual descriptions from sequences of images,(e.g. videos).

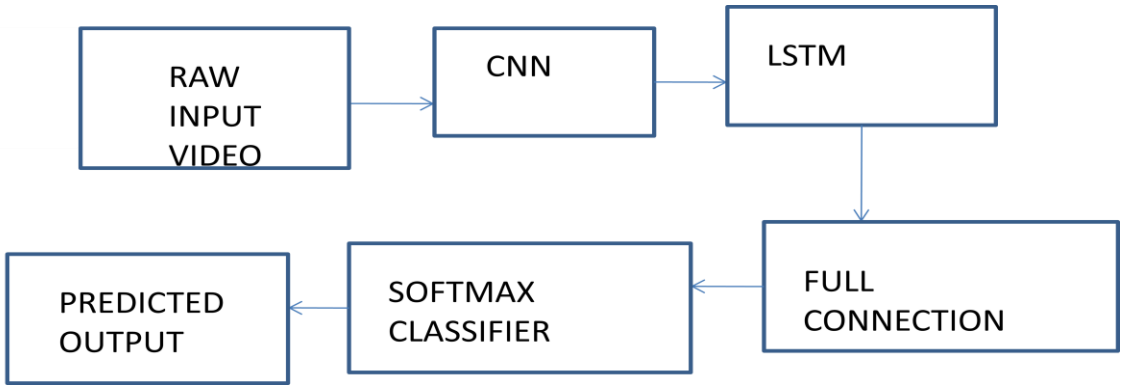

Figure 1: Block Diagram of proposed Method

The UCF101 is an action recognition dataset of realistic action videos, collected from youtube,having 101 action categories. It consists of 13,320 videos of 101 action categories. Split the dataset into train and test. (From UCF101 website) and maintain a CSV file for holding information to extract frames.Then Sub-sampled to 40 frames for each video and it is pass to Inception V3 model. Inception V3 model is a pre-trained model on image-net dataset. Processing blocks are Convolution Layer, Max Pooling, ReLU-Layer (for thresholding) ,Flattening and pass it to Recurrent Neural Network. LSTM model provided by Keras and passed features to the LSTM model and added a few more layers.Finally a layer with soft max activation gives the predicted class.

IV. DATASET DETAILS

The dataset utilized in building our model is UCF101. The UCF101 dataset recordings are gathered from YouTube, having 101 activity classes. UCF101 dataset contains 101 categories which is extension of UCF 50 Dataset. With 13320 recordings from 101 activity classes, UCF101 gives the biggest assorted variety in wording of activities and with the nearness of huge varieties in camera movement, object appearance and present, object scale, perspective, jumbled foundation, enlightenment conditions, and so forth, it is the most testing informational index to date. As the greater part of the accessible activity acknowledgment informational indexes are not sensible and are organized by on-screen characters, UCF101 plans to energize further examination vigorously acknowledgment by learning and investigating new sensible activity classifications.The recordings in 101 activity classes are assembled into 25 gatherings, where each gathering can comprise of 4-7 recordings of an activity. The recordings from a similar gathering may share some normal highlights, for example, comparative foundation, comparative perspective, and so on.

UCF101 incorporates absolute number of 101 activity classes which is isolated into five sorts:Human-Object Interaction,Body –Motion,Human-Human Interaction,Playing Musical Instruments,Sports.

V.IMPLEMENTATION

Implementation steps are executed in the software Pycharm. By using a pre-trained model provided by Keras named Inception-v3 model to extract features from the images. Inception V3 is a model given by Keras whose loads pre-prepared on ImageNet dataset, which is an enormous dataset comprising of huge number of images of various classes. Therefore, in our venture we utilized this Inception-V3 model to extract the features . The LSTM which is a unique sort of Recurrent Neural Network.It is additionally given by keras in python. The feature extraction from Inception-V3, are passed to this LSTM design. From that point forward, we characterized our layers of CNN. We attempted with thick layers with 'relu' initiation with a few dropouts between various layers. In the wake of attempting with various layers, we settled down to a 2048-wide LSTM layer followed by a 512 thick layer and utilized 0.5 dropout. Note that we utilized the highlights from a video as contribution to the LSTM. Rather than passing highlights from each outline, we diminished the quantity of edges to be passed. The features from approximately 30-50 edges for each video are sent as grouping to the LSTM. At the last layer, we utilized 'softmax' activation layer to predict the action recognition.

VI.RESULT AND DISCUSSION

The trained Inception v3 model and CNN-LSTM is tested on UCF101 Dataset. The below outputs are predicted the action recognition by using a proposed model, Figure:2 Apply Eye Makeup, Figure:3 Apply lipstick, Figure:4 Babycrawling, Figure:5 Balance Beam, Figure:6 Band Marching, Figure:7 Baseball Pitch, Figure:8 Basketball , Figure:9 Baseket ball Dunck, Figure:10 Cricket shot, Figure:11 Cricket Bowling, Figure:12 Diving , Figure:13 Drumming, Figure:14 Fencing, Figure:15 Front Crawl.

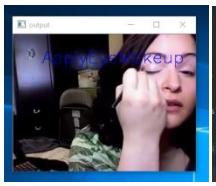

Figure:2

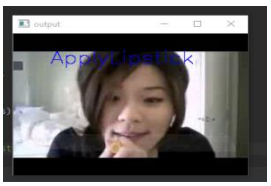

Figure:3

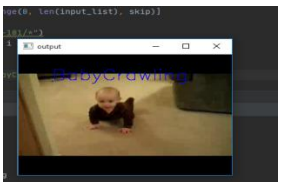

Figure:4

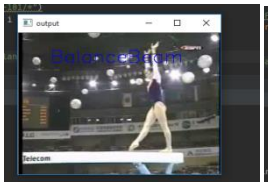

Figure:5

Figure:6


Figure:7

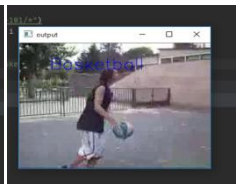

Figure:8

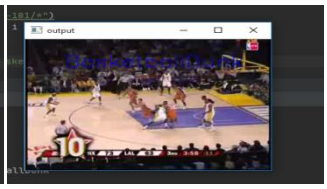
Figure:9

Figure:10

Figure:11

Figure:12

Figure:13

Figure:14

Figure:15

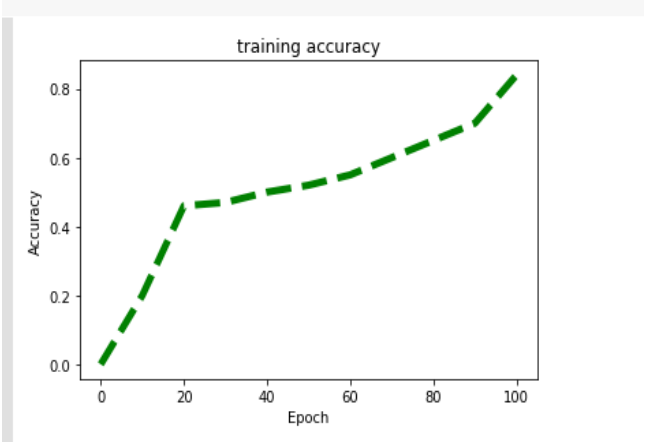

Figure 15: Training Accuracy

Figure 15:shows training accuracy has reached upto 84 percent and that accuracy was achieved at 100 epoch.

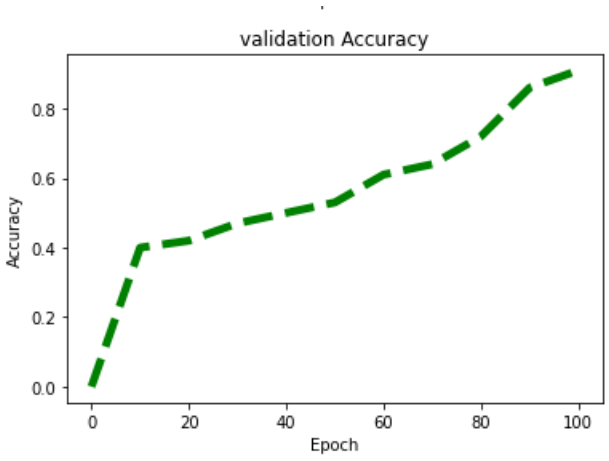

Figure 16: Validation Accuracy

.Fig 16 shows the validation Accuracy for UCF101 data was around 91 percent and was achieved with 100 epoch.

Table 1: Performance comparison of Video classification

METHOD	ACCURACY
Deep Network	65.4
Spatial Stream Network	72.6
LRCN	71.1
LSTM Composite Model	75.8
Relu+CNN	82.3
Proposed Method CNN+LSTM	91

Table 1: shows the performance comparison of Video Classification proposed method CNN-LSTM achieves 91% that is 25.6 higher than Deep network.In addition compared with spatial stream network and Long-term Recurrent convolution netwotk(LRCN) our architecture improves 18.4 and 19.9.Then compare with LSTM composite model and Relu-CNN it improves upto 15.2 and 8.7 respectively.

VI.CONCLUSION

The deep learning is used to solve the human action recognition.The new model used for implementation gave about 91 percent validation accuracy on such big dataset UCF101 .In this paper we have investigated the ability for CNN to learn features from video frames. In future work the Spatio-temporal will be proposed to provide a better way to human action recognition.

REFERENCES

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

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

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

[4] Yann Lecun,Leon Bottou,Yoshua Bengio,and Patrick Haffner,"Gradient-based learning applied to document Recognition," *Proceedings of the IEEE,* pp.2278-2324,1998.

[5] Yuanyuan Huang, Haomiao Yang,and Ping Huang, "Action recognition using hog feature in different resolution video sequences," *International Conference on Computer Distributed Control and Intelligent Enviromental Monitoring*, 2012.

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

[7] Georgios Th. Papadopoulos, Apostolos Axenopoulos and Petros Daras,"Real-time skeleton-tracking-based human action recognition using kinect data," *Proceedings of the International Conference* ,pp:0302-9743, 2014.

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

[9] Samitha Herath, Mehrtash Harandi,and Fatih Porikli,"Going deeper into action recognition,"*A survey of Image Vision Computation*, 2017.

[10] Jindong Wang et al.,"Deep learning for sensor-based activity recognition,"*A Survey Pattern Recognition Letters Elsevier*,pp.1-10,2018 .