ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse199942020.pdf
https://doi.org/10.30534/ijatcse/2020/199942020

Transfer Learning Based Activity Recognition using ResNet 101 C-RNN Model

Neha Mangal[1], Aaditya Jain[2]
[1]Rajasthan Technical University, Kota, Rajasthan, India
[1]14neha10@gmail.com
[2]Faculty of Engineering and Computing Sciences, Teerthanker Mahaveer University, Moradabad, India
[2]:aadityajain58@gmail.com

ABSTRACT

As we know traditional passive video surveillance systems are inefficient and error prone, as they need human eye for analysis of recorded data. Automatic Human Activity Recognition is currently receiving an increasing consideration from computer vision scientists with the goal of an automated analysis of ongoing activities from video streams. This interest is inspired by its applications over a widespread field such as smart surveillance systems, border crossings, health care systems, robot learning etc. Deep networks allow building such intelligent surveillance systems with automated feature learning. But large volume of data is required for sufficient network training, which in-turn demands for more computational resources to conduct network training. There is strong need for solutions that can better exploit available data and require minimal training and preprocessing. This paper investigates the use of the CNN and ResNet 101 model known as C-RNN architecture for activity recognition from video streams using transfer learning. The proposed approach is prepared and approved on well-known UCF-101 dataset based on accuracy and average loss which gives preferred outcomes over state-of-art methods. This approach gives impressive performance and great potential for computer vision.

Key words : Deep Learning Approaches, Transfer learning, UCF 101, ResNet 101, Activity Recognition.

1. INTRODUCTION

We are in the midst of a data revolution where visual content plays a protagonist role. Around 90% of data that comes from the human brain is visual, and our brain is capable of processing visual information 60000 times faster than text. The exponential growth of portable video cameras and online multimedia repositories, as well as recent advances in video coding, storage and computational resources have motivated intense research in the field towards new and more efficient solutions for organizing, understanding and retrieving video content. Video systems are employed for monitoring inside and outside of businesses, offices and remote locations are not mere for recording of data. Therefore, there is strong need of an active surveillance system that can automatically detect abnormal or suspicious activity.

Expanding prominence of such wise frameworks is essential because of expanding applications in various areas like surveillance, human fall location, healthcare or in ambient assisted living, computer human interaction, and furthermore for games examination purposes. Numerous private residents are deciding to introduce home camera reconnaissance frameworks so as to screen their own friends and family, either in age-related consideration offices or at home. Automatic surveillance systems involve the transmission signal to a concerned specialist about any suspicious exercises. Also, ambient assisted rooms, permits constant observing of exercises of patients to doctors and nurses. Acknowledgment of exercises from the video stream includes learning action portrayal or highlights followed by their arrangement. Recognition of human activities is a contemporary keynote in the field of computer vision. As of now human activities are recognized under two categories one consists of hand crafted and another is learning based.

1.1 Handcrafted Based Approach

The most traditional approach which is popular among the HAR community from past decades and shown very interesting results among well-known datasets is known as handcrafted based approach. In this approach, comprehensive features are collected from the sequence of images and their classification is done by using a classifier like support vector machine. Because of the progression in sensor and visual innovation, HAR based frameworks have been broadly utilized in some real time applications. Specifically, the proliferation of compact sized sensors has empowered the intelligent gadgets to recognize the activities of humans in a context-aware manner.

1.2 Learning Based Approach

Dissimilar to handcrafted based methodology, another methodology for human movement acknowledgment is Learning based methodology. In the learning based methodology, the highlights are found out consequently lessening the relentless human intercession, master information and choice of ideal highlights. The major challenge for learning models is voluminous video datasets for preparing purposes and training time.

Deep learning techniques have recently become the novel state-of the art in many computer vision tasks and mainly focuses on layered architecture. These approaches are further categorized into supervised learning as well as non-supervised learning. In supervised learning labels are used for training purposes while in unsupervised learning no labels are used for training purposes. Deep learning includes automated learning of features using CNNs and RNN's. Especially CNN is used for activity recognition purposes because of its self-learning capacity without any need of external modality. In this paper, we discourse the problematic of automatic recognition of human activities from videos. Our goal is to develop 3D CNN and ResNet 101 C-RNN based system that can automatically learn activities from provided set of training videos, and recognize them in unseen, diverse, and realistic videos [1][2].

Section 1 discusses about the introduction, section 2 focuses on related work while Section3 presents proposed methodology in detail. Section 4 describes dataset used, experimental settings. Section-5 includes results, discussion and comparison with previously proposed works. Section-6 concludes the work and also discusses avenues for future research based on the proposed approach.

2. RELATED WORK

Recognition of human activities can be performed at various degrees of reflection. Throughout the latest years, diverse logical orders have been proposed to describe these degrees of reflection [3]. In this segment we will contemplate the various methodologies dependent on profound learning.

Guha *et al.*[4] presented an approach in which 2D CNN is extended to 3D CNN yielding a fully automated deep learning based framework having RNN as classifier for each learned spatio temporal sequence of each time step. Yongmou *et al.*[5] presented the strategy of extracting hand crafted features by using gyroscope and accelerometer sensor data. Unsupervised feature learning by using PCA, sparse auto encoder makes the framework to learn features automatically from massive training data.

Ryoo *et al.*[6] compared the available methods of learning convolutional neural networks such as pooling, image based CNN and RNN's on robot centric videos having robot human interactions or videos in the outdoor regions. Basura *et al.*[7] proposed a rank pooling machine in which Temporal ordering of the video is preserved by training the linear ranking machine according to ranking of the frames capturing appearance and evolution over time by supervised learning. Earnest e*t al.*[8] presented the approach in which action banks extracted in two ways are used in the convolutional neural network as input. Action banks are used in extracting the feature by convolutional neural network.

Joe et al.[9] presented a method in which spatio temporal features from long term video frames are learned over LSTM using GoogleNet and AlexNet Models. Along with them RGB and Optical flow are used as the source of input for recognition purposes and feature pooling is used to classify actions. After the huge success of CNNs over images these are extended for 1 million videos Sports 1M dataset [10] having huge variation of activity classes approximately 487. In addition, 2 spatial resolutions one is low resolution context stream and other one is high-resolution fovea streams are combined to attain outstanding results and lessen training time. To validate the results of other interesting data sets, the perception of transfer learning was used for data set UCF 101 and slow fusion networks, and better outcomes were achieved in only a few layers of fine-tuning, rather than from scratch learning.

Recently Cheng et al. [11] presented the approach in which Body Activity Recognition using data from wearable sensors. Classifying of the actions is done by using different approaches such Machine learning algorithms artificial neural network. Some of the recent works are contributed by [12] in which dynamic images are constructed using Dynamic Depth, Dynamic Depth Normal and Dynamic Depth Motion Normal. Then these dynamic images are used for further training and classification purposes.

Rajat et al. [13] proposed a four stream model in which first stream consists of training of the pre trained VGG net with dynamic images and other three streams are trained via front depth, side and the top motion map on the VGG net. Dynamic images having information of full video into one image are constructed using rank pooling from a video. Khurana et al.[14] drafted an approach in which there is fusion between spatial stream and spatio temporal streams. In spatial stream 2D filters are used to separate highlights from RGB pictures, whereas in spatio temporal stream 3D filters used to separate highlights from RGB pictures. Both streams are individually trained on pre-trained network and softmax scores are fused for the final activity class.

Mehrjou A et al.[15] combined different vision modalities such as RGB data, skeletal data and depth data from RGB-D sensor. Different input modalities containing images are trained separately using the convolutional model and their fusion score is combined at decision level for action recognition.

Training of deep networks is resource consuming as it requires very high-speed multicore systems, therefore benefitting from pre-trained models is an avenue to explore, which will save time as well as money. Much of research is focused on attaining high accuracy, real time response, and faster analysis [16][17]. Therefore, future research needs to be focused on designing efficient as well as fast video analysis systems. So this paper moves in this direction.

3. PROPOSED APPROACH

Traditional passive video surveillance systems are inefficient and error prone, as they need human eye for analysis of recorded data. Deep networks allow building such intelligent surveillance systems with automated feature learning. But the large volume of data is required for sufficient network training, which in-turn demands for more computational resources to conduct network training. There is strong need for solutions that can better exploit available data and require minimal training and preprocessing.

A video comprises of arrangements of pictures or edges along the worldly measurement. Distinguishing proof of movement can be essentially cultivated by utilizing 2D convolutions on pictures/outlines independently to learn action portrayal. This methodology doesn't consider movement encoded in outlines. ID of certain exercises is conceivable by utilizing a static casing however doesn't remain constant for different exercises. Thus, various methodologies are considered for fleeting data. Utilizing some extra info methodology like optical stream, dynamic pictures and twofold movement pictures and so forth is one approach to do this. Commitment of extra info methodology to learn action class names can't be overlooked, and yet it needs extra pre-preparing of video information to get wanted info methodology. Accordingly, here, work thought is to get tantamount outcomes utilizing just RGB outlines and insignificant preparing by utilizing profound leftover models as opposed to shallow systems. In the proposed ResNet 101 C-RNN model we have consolidated highlights of CNN too of RNN. To validate our research, we have proposed two variety of model i.e. one using only 3D CNN model and one having CNN as well as RNN.

3.1 Overview of 3D CNN Model

In the 3D CNN model we have proposed an approach to fetch capabilities of CNN alone. The proposed solution comprises of 3D input from videos. Input video is preprocessed to 90x120 size frames. Then these RGB frames are feed into the 3D CNN Model, which consists of 2 Convolution layers, two polling layers and 2 fully connected layers having size of 2048 neurons. ConV_1 is having a size of 64x5x5 Conv2 with 128x3x3 both having stride of 2. Softmax classifier is used for calculating prediction scores for particular activity. The architecture for above stated model is shown in figure 1 below. Model is trained on UCF 101 dataset by using three train-test splits. Proposed 3D CNN model shows in figure 1.

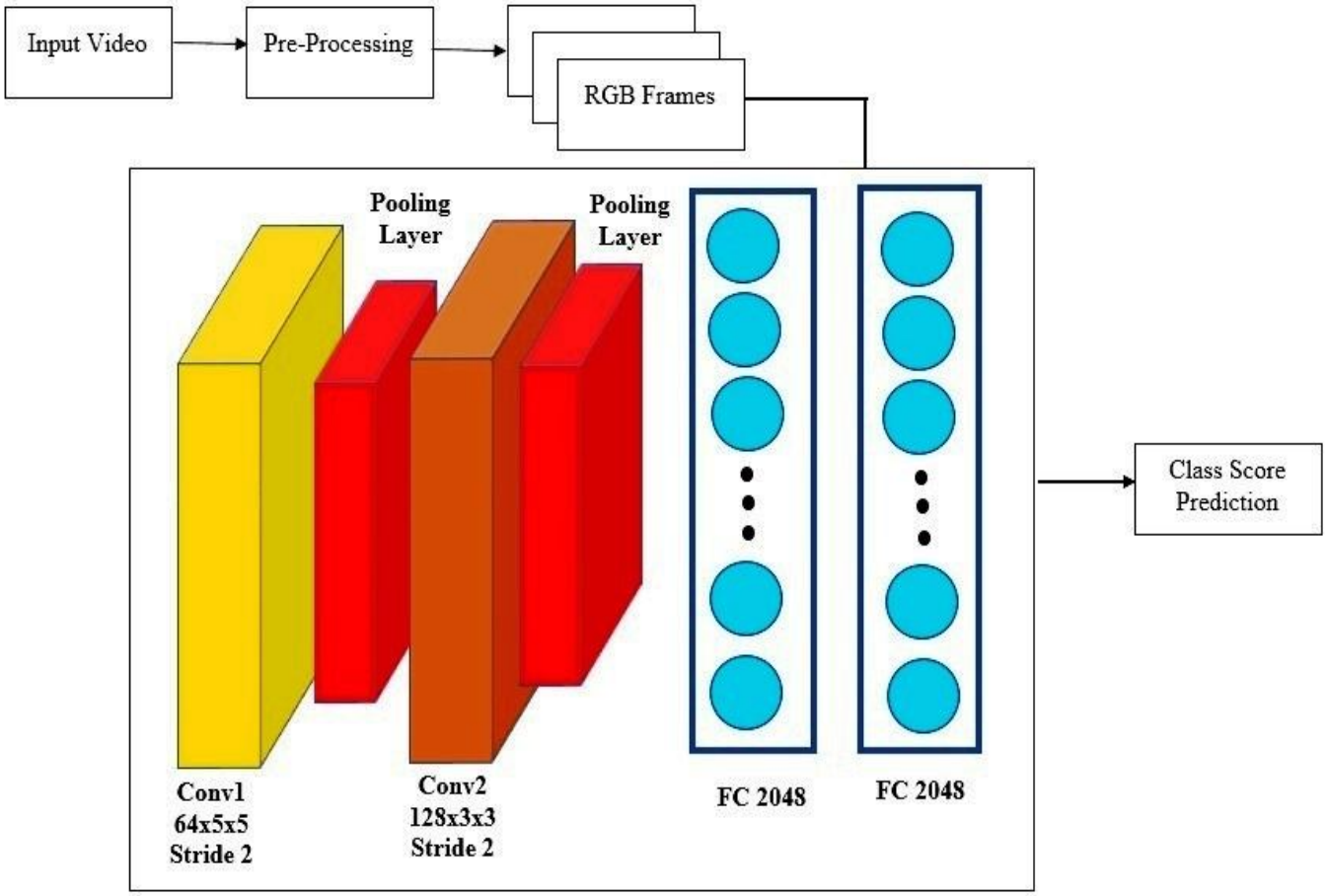

Figure 1: Proposed 3D-CNN Model for Human Activity Recognition

3.2 Architecture of ResNet 101 C-RNN Model

In this proposed ResNet 101 C-RNN model we have fused features of the Convolutional neural network as well as of Recurrent Neural Network. The CNN architecture acts as encoder whereas ResNet101 C-RNN acts as a decoder. In the process of encoder, Two-dimensional CNN architecture is deployed having four layers and three Fully connected layers. Four convolutional layers, namely ConV_1, ConV_2, ConV_3, and ConV_4 having size of 32x32, 64x64, 128x128 and 256x256 respectively. Three fully connected layers having dropout of 0.3 is used. The encoded CNN encodes each 2D picture X(t) into a 1D vector Z(t) by utilizing condition 1. CNN goes about as encoder which encodes all the data produced from input pictures into a solitary vector.

$$f_{\mathrm{CNN}}(\mathbf{x}^{(t)}) = \mathbf{z}^{(t)} \quad \text{..... (1)}$$

Then the decoder comes into action which uses this vector as input and trained on ResNet 101. ResNet 101 is trained on ImageNet by using transfer learning scheme. The capabilities of ResNet 101 are transferred to our Encoded data. This consequently predicts the action class. The design of the proposed C-RNN ResNet 101 model is shown in figure 2.

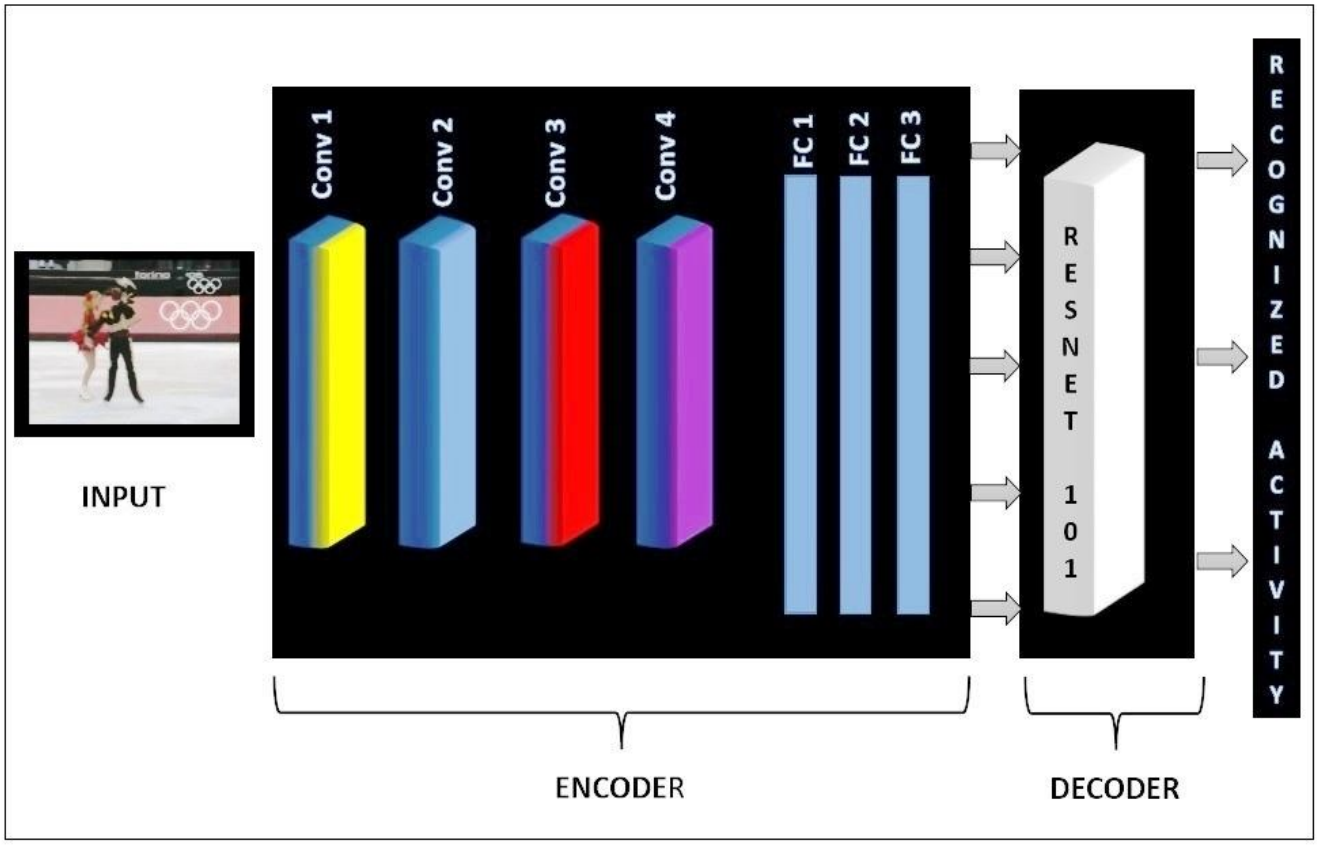

Figure 2: Our proposed C-RNN ResNet 101 Architecture

3.3 Algorithm for C-RNN 101 Model

Step1. Start
Step 2. Apply pre-processing which acts as input.
Step 3.Extract frames from the UCF 101 dataset. And prepare images of size (90 x 120) for each action sequence.
Step 4. Train the Model on C-RNN as proposed in Section 3.2 for input with the network architecture discussed in Fig 2 by using equation 1. The encoded information from CNN is feed into RNN for decoding purpose.
Step 5. Output generated from RNN effectively decides the class of activity.
Step 6. END

4. EXPERIMENTS

Hardware and software setup used in experiments is given in section 4.1. Experiments are performed on widely used and challenging benchmarks for action recognition: UCF-101 described in section 4.2. Section 4.3 specifies training and testing setup.

4.1 Hardware Setup & Software setup

Figure 3 focuses on hardware and software settings used for experiments. The network's size is limited by resource constraints mainly the amount of GPU memory and training time that one is willing to tolerate. Hardware is composed of CPU and Nvidia Quad GPU having 8 gigabytes of memory. Linux OS Ubuntu 16.04 is used in conjunction with Python and CUDA toolkit. Python provides various deep learning libraries and CUDA toolkit supports GPU-accelerated computing. PyTorch deep learning development platform is used, that provide very simple API for implementation of neural nets.

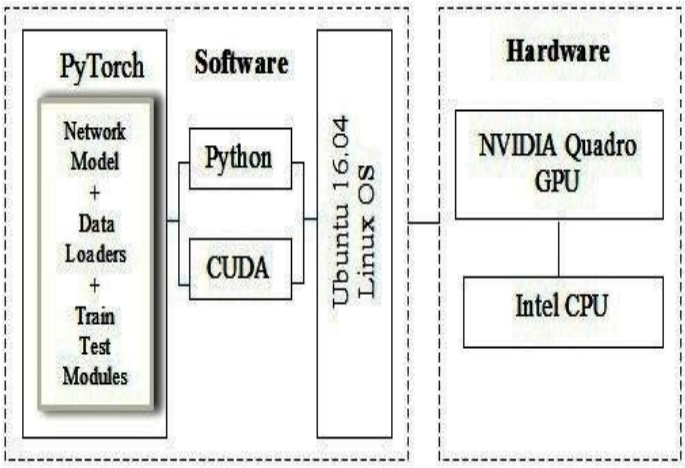

Figure 3: Hardware and software setup

4.2 Dataset Used

UCF101 [18] is commonly used video dataset that consists of total 13320 video. Basically total 25 groups were formed to classify different action videos where each group has 5 to 8 video of certain action.

An action can be categorized by five ways: 1) Body-Motion Only 2) Human- Object Interaction 3) Human-Human Interaction 4) Playing Musical Instruments 5) Sports. Fig.4 represents frames of some four different activity classes. The cause behind choosing this dataset is that it offers the high variety of posture, camera gesture, object appearance, scale, and lighting conditions. This allows testing and verifying the robustness and effectiveness of the recognition model in the harsh real-world scenarios.

Brushing teeth Knitting Drumming Horse riding

Figure 4: UCF 101 dataset Samples

4.3 Network Training and Testing

The first step for network training consists of extraction of RGB frames for each dataset using ffmpeg. Mini batch training is used, that is the combination of batch and stochastic training, as it uses a specified number of items (batch size) to compute gradients. Batch size is adjusted to fit data in available GPU memory. 29 frames per videos are used for training. Mini batch having size of 30 is used, that is the combination of batch and stochastic training, as it uses a specified number of items (batch size) to compute gradients. Also, the initial learning rate of 1e-4 is used. With experimentation we tried to adjust batch size to fit data in available GPU memory. Transformation functions are applied to video frames for data augmentation and generalization of a trained model. We have run 120 iterations for our proposed model and calculated validation Accuracy as well as loss by using 19 frames.

5. RESULTS AND DISCUSSION

Segment 4 talks about the exhibition estimates, for example, Accuracy and loss during the validation stage. Additionally, our proposed strategy is contrasted and best in class strategies based on accuracy. Section 5.1 discusses about the average loss and accuracy of the 3D CNN model as proposed in Section 3.1. Section 5.2 discusses about average accuracy and loss for the ResNet 101 C-RNN model as proposed in Section 3.2. Section 5.3 discusses about comparison of our proposed method with other state-of-art methods.

5.1 Results: 3D CNN Model

Accuracy and average loss for the 3D CNN model is shown in Fig.3 and Fig 4 respectively for 120 epochs. The Validation accuracy of 3D CNN is 55.89% with average validation loss as 2.81%. The results shown in figure 5 and 6 and it proves that alone CNN is not able to classify the activities with good accuracy.

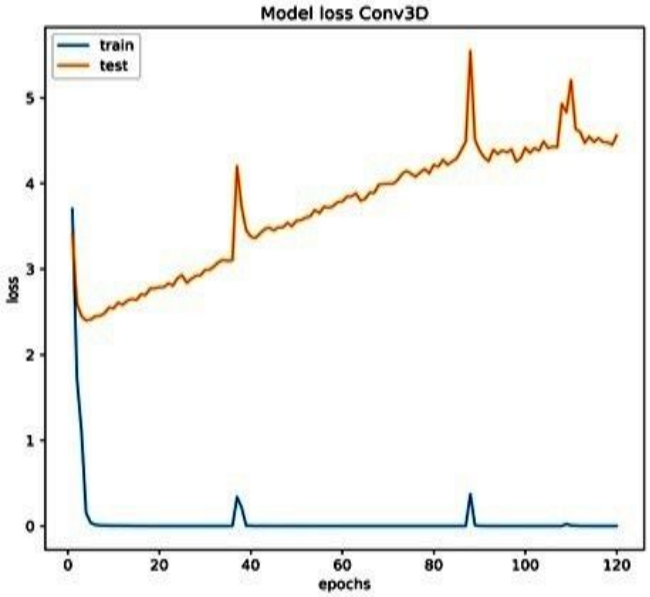

Figure 5: Average Loss of 3D CNN Model

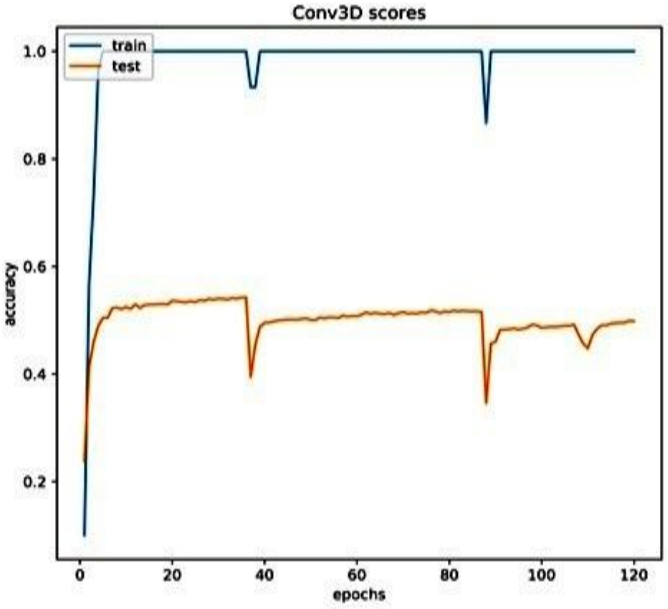

Figure 6: Accuracy of 3D CNN Model. Best Epoch 35

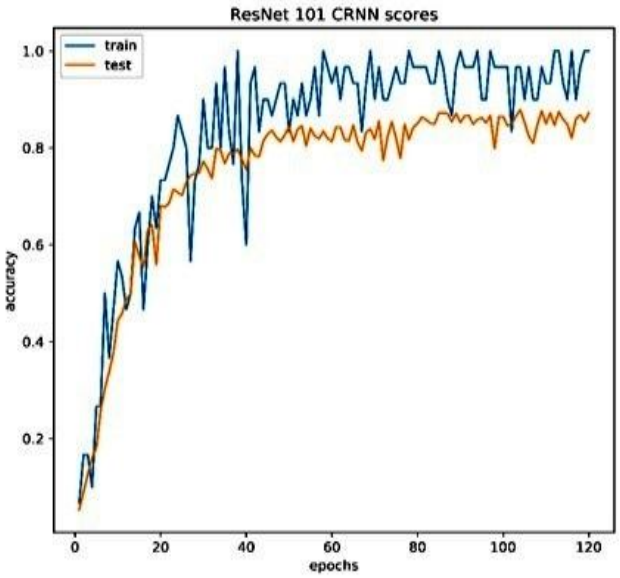

Figure 8: Accuracy for ResNet 101 C-RNN Model. Best Epoch 87.

5.2 Results: ResNet 101 C-RNN Model

Accuracy and average loss for ResNet 101 C-RNN model is shown in Fig.7 and 8 respectively for 120 epochs.Approval precision of ResNet CRNN is 91.04 % with normal approval loss of 0.95%. Outcomes show that there is increment of precision just about 35% from CNN model. As we have utilized ResNet 101 pre prepared model which increment the precision as appeared in figure 7 below.

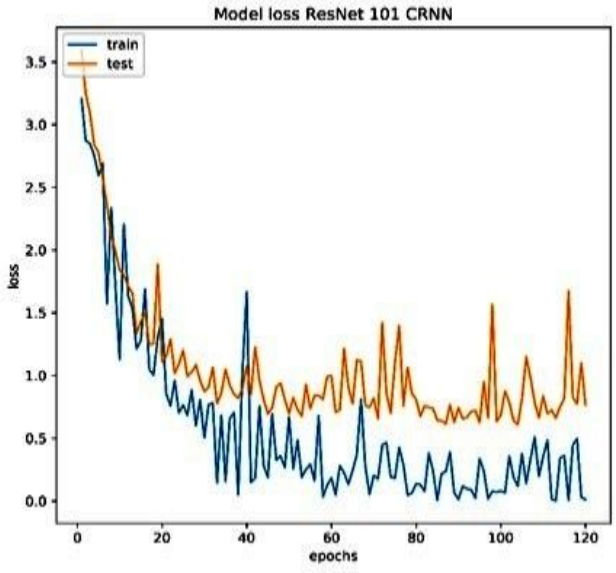

Figure 7: Average Loss of ResNet 101 C-RNN

5.3 Comparison with State of Art

The method proposed above ResNet 101 C-RNN uses only RGB frames. Hence, to support our research we have compared our models which mainly use RGB frames trained on UCF-101. Table.1 shows the proposed model with other methoods. Results for3D CNN model (proposed) and ResNet 101 C-RNN (proposed) outperforms spatial stream (VGGM-2048) in [7][12]by good difference.

Table 1. Comparison of proposed method and existing approaches.

Sr. No.	Method	Input modality	Recognition Accuracy (%)
1	[4] slow fusion	RGB	65.4
2	[12] spatial stream	RGB	72.7
3	[7] spatial stream	RGB	74.2 (VGGM-2048)
4	[7] Temporal stream	Optical flow	82.34 (VGGM-2048)
5	[13] Decision fusion	Dynamic flow and RGB	84.93
6	[13] Fusion using SVM	Dynamic flow, RGB, optical flow	88.63
7	3D CNN Model	RGB	**55.89**
8	ResNet 101 C-RNN	RGB	**91.04**

6. CONCLUSION

This paper proposed two models namely 3D CNN model and ResNet 101 C-RNN model. The results of residual C-RNN clearly highlights that deeper nets having residual connections have the ability to contribute to notable progress in fields related to various video analysis tasks. The results

obtained are comparable to state-of-the-art models that have even used additional input modality. As future work, we will explore performance with some other modality such as handcrafted features are combined with 3D residual networks for recognition of activities and may also consider other standard video dataset.

REFERENCES

[1] Jain A, Soni B, "Secure Modern Healthcare System Based on Internet of Things and Secret Sharing of IOT Healthcare Data", International Journal of Advanced Networking & Applications (IJANA), Vol. 8, Issue 6, pp:3283-3289, ISSN: 0975-0282, 2017.

[2] Jain A, Soni S, "Visual Cryptography and Image Processing Based Approach for Secure Transactions in Banking Sector", 2nd **IEEE** International Conference on Telecommunication & Networks (TELNET), PP. 01-05, August 2017, DOI: 10.1109/TEL-NET.2017.8343545.

[3] Jain A, Mangal N, "Deep Learning Approaches for Activity Recognition", International Journal of Control and Automation, 13(4), pp. 919-927, 2020.

[4] Guha T, Member S, Ward RK. Sequential deep learning for human action recognition. Hum Behav Underst. 2011;29--39.

[5] Li Y, Shi D, Ding B, Liu D. Unsupervised Feature Learning for Human Activity Recognition Using Smartphone Sensors. Expert Syst Appl. 2014;41(14):6067–74. https://doi.org/10.1016/j.eswa.2014.04.037

[6] Ryoo MS, Matthies L. Video-based convolutional neural networks for activity recognition from robot-centric videos. 2016;9837:98370R. Available from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2229531

[7] Fernando B, Gavves E, Jose Oramas M, Ghodrati A, Tuytelaars T. Rank Pooling for Action Recognition. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):773–87.

[8] Ijjina EP, Mohan CK. Human action recognition using action bank features and convolutional neural networks. 2014 Asian Conf Comput Vis [Internet]. 2014;59:178–82. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7033111

[9] Ng JYH, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: Deep networks for video classification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2015;07–12–June:4694–702.

[10] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li FF. Large-scale video classification with convolutional neural networks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2014;1725–32. https://doi.org/10.1109/CVPR.2014.223

[11] Cheng L, Guan Y, Kecheng Zhu, Yiyang Li. Recognition of human activities using machine learning methods with wearable sensors. 2017 IEEE 7th Annu Comput Commun Work Conf [Internet]. 2017;1–7. Available from: http://ieeexplore.ieee.org/document/7868369/

[12] Wang P, Li W, Gao Z, Tang C, Ogunbona PO. Depth Pooling Based Large-Scale 3-D Action Recognition with Convolutional Neural Networks. IEEE Trans Multimed. 2018;20(5):1051–61.

[13] Singh R, Khurana R, Kumar A, Kushwaha S, Srivastava R. Combining CNN streams of dynamic image and depth data for action recognition. Multimed Syst [Internet]. 2020;(0123456789). Available from: https://doi.org/10.1007/s00530-019-00645-5

[14] Khurana R, Kumar A, Kushwaha S. Delving Deeper with Dual-Stream CNN for Activity Recognition [Internet]. Springer Singapore; pp. 333-342 Available from: http://dx.doi.org/10.1007/978-981-13-2685-1_32

[15] Mehrjou A, Hosseini R, Araabi BN. Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognit Lett [Internet]. 2018;2–8. Available from: http://dx.doi.org/10.1016/j.patrec.2015.10.004

[16] Jain A, "Clustering of Text Streams via Facility Location and Spherical K-means", IEEE International Conference on Electronics, Communication and Aerospace Technology (ICECA-2018), pp. 1209-1213, 2018. DOI: 10.1109/ICECA.2018.8474757.

[17] Jain A, Tyagi S, "Priority Based New Approach for Correlation Clustering", IJITCS, MECS Press, Vol. 9, Issue 3, pp. 71-79, 2017, DOI: 10.5815/ijitcs.2017.03.08.

[18] Soomro K, Zamir AR, Shah M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. 2012;(November). Available from: http://arxiv.org/abs/1212.0402

[19] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos. Adv neural Inf Process Syst. 2014;568–76.

[20] Feichtenhofer C, Pinz A, Zisserman A. Convolutional Two-Stream Network Fusion for Video Action Recognition. 2016;(i). Available from: http://arxiv.org/abs/1604.06573

[21] Wang J, Cherian A, Porikli F. Ordered pooling of optical flow sequences for action recognition. Proc - 2017 IEEE Winter Conf Appl Comput Vision, WACV 2017. 2017;168–76. https://doi.org/10.1109/WACV.2017.26