2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS) | 978-1-7281-7136-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ACCESS51619.2021.9563316

Real-time Human Activity Recognition Using ResNet and 3D Convolutional Neural Networks

Archana N.
Dept. of Electronics and Communication Engineering
Government College of Engineering Kannur
Kannur, India
archanavazhayil@gmail.com

Hareesh K.
Dept. of Electronics and Communication Engineering
Government College of Engineering kannur
Kannur, India
hareesh@gcek.ac.in

***Abstract*—In computer vision-based applications, the recognition of human activity is always a standard problem. Nowadays, activity recognition is more possible and accurate due to good development in artificial neural networks like convolutional neural network CNN. In many recent works, the recognition model architecture use CNN and long short-term memory units (LSTM) - attention models to extract spatial and temporal features from the input video. This particular work is related to real-time human activity recognition by Resnet and 3D CNN without the involvement of the LSTM- attention model. Here the 2D Resnet is modified to 3D CNN to achieve better human activity recognition accuracy. The wide range of data information from the kinetics dataset can avoid overfitting issues during the training period. And the combination of Resnet and 3D CNN can enhance the accuracy of recognition. As a consequence, a method for detecting, monitoring, and recognizing real-time human motion has been developed.**

***Index Terms*—Real-time human activity recognition, Spatio-temporal features, 3D Convolutional neural networks, Resnet-18, kinetics 400 datasets**

I. Introduction

Human activity recognition is a major part of some computer vision-based applications. There are so many methods to achieve human activity recognition [1]-[2][3][4]. The main consideration during human activity recognition is feature extraction. In an input video clip, both spatial and temporal features are extracted for activity recognition [1]. In some of the cases, the apparent motion of brightness pattern is considered as a feature for identification of motion, also known as optical flow [5]. The recent growth in the area of artificial neural networks like CNN and deep learning [6][7] techniques made human activity recognition more easy and accurate. At the same time occlusion, camera motion, insufficient datasets to train the network, etc. are making recognition more complicated.

Action classification and action identification are used to recognize human activity. Action representation methods and interaction recognition methods are included in action classification. Some of the feature descriptors used to describe motion include a motion boundary histogram and an optical flow histogram [5].This feature is often used to describe and extract the video's spatial and temporal shifts. The methods for portraying similar features in RGB data are based on skeleton joint trajectory, human image series, spatiotemporal volume, and so on. The depth and skelton data handcrafted action feature representation are based on three methods: depth sequence-based method, skeleton-based method, and feature fusion method [8]. For action feature representation, deep learning techniques along with two-stream convolution networks, 3D convolution networks, and a special recurrent neural network called LSTM are now being used for human activity recognition. Confusion matrix and accuracy are often used as criteria for assessing recognition results. Deep learning approaches outperform current methods in terms of accuracy and efficiency. For human activity identification, the HMDM51 and UCF101 datasets are usually used [9]. However, it may also show a lack of ability to train the network, resulting in overfitting.

Human behavior detection is a common concern in machine vision-based systems. Thanks to the advancement of artificial neural networks such as the convolutional neural network (CNN), human behavior detection is now more realistic and precise. In most cases, the recognition model architecture extracts spatial and temporal features from the input video through CNN and long short-term memory units (LSTM)-attention models. This work focuses on real-time human activity detection using only Resnet [10] and 3D CNN [11], without the use of the LSTM-Attention model [1]. To improve human action detection accuracy, here the 2D Resnet is converted to a 3D CNN. The kinetics dataset's large variety of data can help to prevent overfitting during the training phase. Besides, combined Resnet and 3D CNN can improve the recognition accuracy.

The rest of this paper is laid out as follows: In Section 2, the paper goes through the 3D CNN and Resnet-18 models, as well as preparation and research strategies. Then, in Section 3, present the experimental results on the kinetics dataset. Finally, Section 4, bring the study to a close.

II. Proposed 3D Convolution Neural Networks and ResNet

CNNs (convolutional neural organizations) are a type of profound neural organization model that can work on crude data sources straightforwardly. However, such models can only handle 2D inputs at a given moment. By performing 3D

978-1-7281-7136-4/21/$31.00 ©2021 IEEE

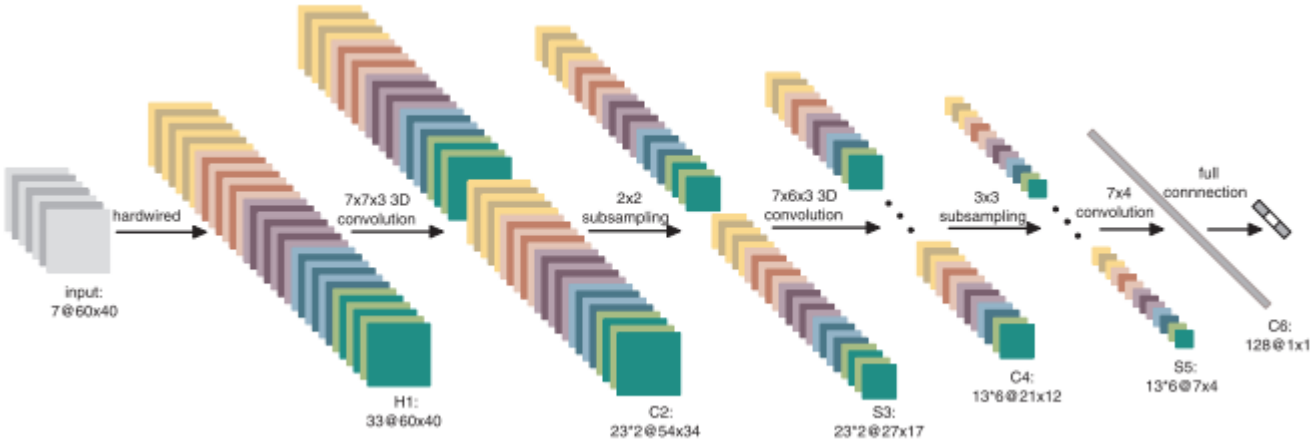

Fig. 1. A 3D CNN architecture for activity recognition.[11]

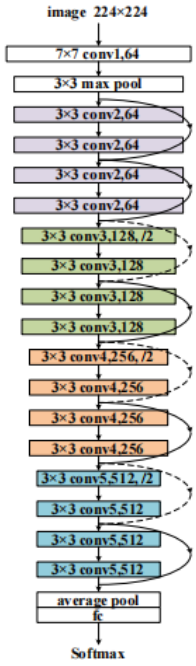

Fig. 2. Network structure of ResNet-18.[12]

convolutions, the 3D convolutional neural organization model gets qualities from both the spatial and fleeting perspectives, catching the movement data encoded in a few contiguous frames. From the info outlines, the 3D CNN model creates a few channels of data, and the last component portrayal coordinates data from all channels. To further develop precision, regularization of the yields with significant level elements and coordinating the evaluations of a few distinct models are utilized. A typical 3D CNN architecture is shown in Fig. 1.

ResNet-18 has comparable results to other ResNets, but since it is shallow, it can keep more low-scale functionality. Subsequently, the ResNet-18 pre-prepared model can be utilized as a function extractor for network models. ResNet-18 has 16 convolutional layers, two downsampling layers, and a few entirely linked layers (Fully Connected layers). ResNet's input picture is 224×224 pixels, the primary convolution layer and the convolution filter size is 7×7, and other layers of CNNs are 3×3 in size [13]. After average pooling the last convolution layer's feature map, a complete relation yields an eigenvector, which is then normalized with Softmax to yield the classification likelihood. As seen in Fig. 2, two convolution layers of a comparative concealing design an extra square that yields a comparative size feature map and has the comparative number of filters.

III. EXPERIMENT AND ANALYSIS

A. DATASET(KINETICS 400)

The dataset is primarily concerned with human actions. Person Actions (singular), such as drawing, balloon blowing, and archery; Person-Person Actions, such as massaging, kissing, and shaking hands; and Person-Object Actions, such as folding clothes, playing the keyboard, and washing dishes. Instead of a deep hierarchy, there are several (non-exclusive) parent-child groupings, such as music (keyboard, flute, guitar, etc.); cooking (grinding meat, making pizza, peeling, etc.) [14]. Fig. 3 depicts video clips from a variety of classes.

B. IMPLEMENTATION DETAILS

This subsection describes how to train the network for real-time human activity identification on the wide kinetics dataset. The kinetics dataset contains about 400 distinct models of action categories. Then the whole data from the kinetics dataset sectioned into, 80% data for training and 20% data for testing. This is on the grounds that activity location is more intricate if the item utilized for the preparation interaction is tiny. So always use a significant number of samples for the training process than the testing process. To make more data for the training process, data augmentation is used here. It is helpful to make more different data of the same object with a different orientation, scale, etc. It is not an essential step, an optional one. It will help to make the datasets much wider.

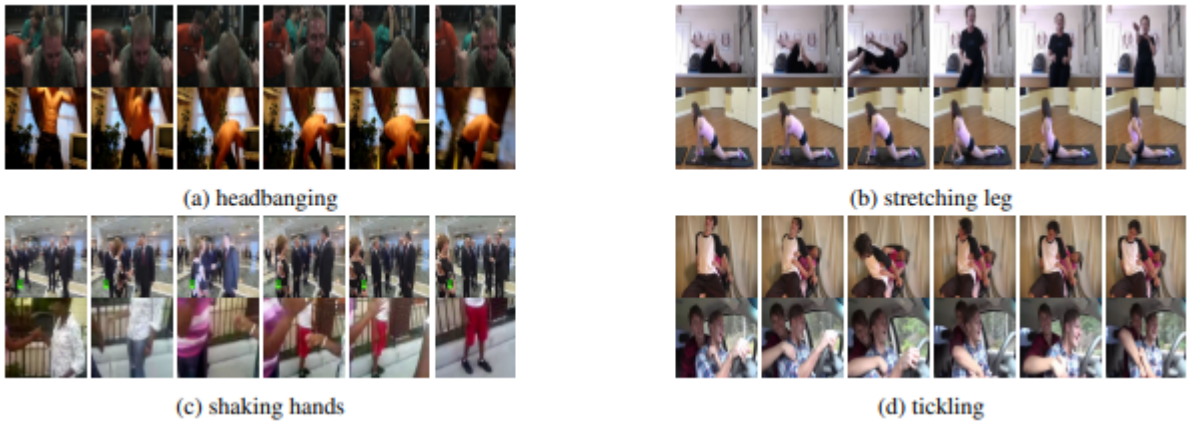

Fig. 3. Kinetics dataset example classes[14]

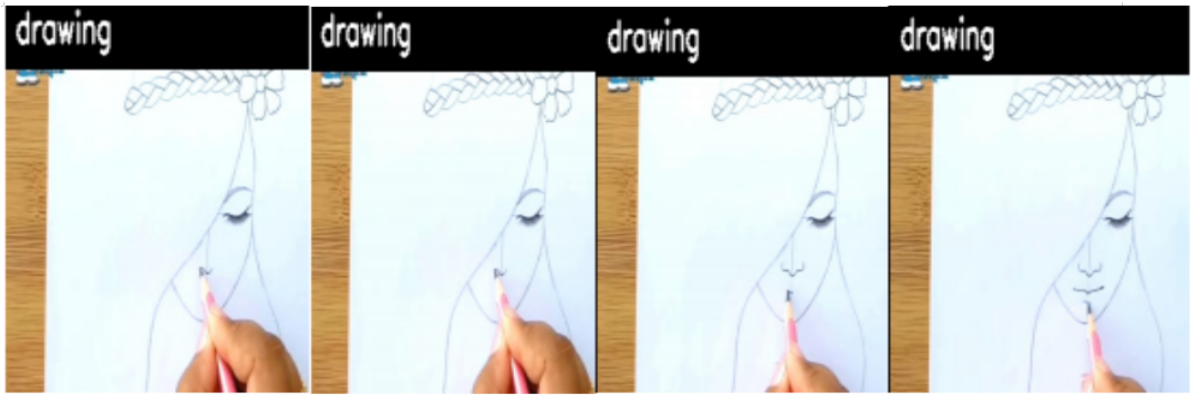

Fig. 4. Examples of successful predictions using the proposed method(Drawing).

Fig. 5. Examples of successful predictions using the proposed method(Balooning).

Fig. 6. Examples of failed predictions using the proposed method.

Here the training set is selected for the augmentation. The normal techniques for augmentation are flipping, zooming, rotating, color jitters, grayscale, etc. Here chose to make 10-degree random rotation and 10 percentage horizontal, vertical flipping. Here the neural networks uses 16 frame clips at the time of training [13]. 3D CNN is capable of extracting spatial and time-related features from raw input. That is, here ResNet-18 network extract both time and appearance-dependent information from the relevant frame for action detection. The advanced gradient descent algorithm is used here to train the network. To take a look over to the gradient vanishing problem, the back propagation algorithm is taken and the cross-entropy is considered as a cost function. The network's rate of learning is fixed as 10^{-3}.

Latter the detection part uses a sliding window system, then the raw video is separated into 16 frames without overlap, then each one is given to the pre-trained network. Then the class score of each frame is found out. The class that has the greatest score demonstrates the distinguished class mark. All tests are implemented based on Tensorflow and Keras.

C. RESULTS AND DISCUSSIONS

The proposed ResNet network is evaluated on the Kinetics 400 dataset in the first step. ResNet's residual learning architecture is adequate for locating spatial knowledge about the frame. This could also prevent gradient vanishing during the backpropagation process. By considering the weights of nodes, the alternative path given by the ResNet-18 network prevents unwanted nodes. The Kinetics 400 dataset is valuable in avoiding overfitting. Fig. 7 shows the accuracy graph for the proposed Model (ResNet-18). This figure shows that increasing the number of iterations reduces both test and training failure. However, over a certain number of iterations (nearly 50 above), the loss becomes saturated. The network's test accuracy often exhibits several specific responses to the number of iterations. The test accuracy is going up at first, however is saturated after reaching a certain iteration threshold (nearly 50 above). Fig. 4 and 5 show the correct predictions for some test frames and Fig. 6 shows the failed one.

IV. CONCLUSION

Human physical behavior identification is a significant step in most computer vision-based technologies. Moreover, real-time motion detection is more useful for real-time applications like the detection of unusual activities of prisoners in jail. Here, the system that is employed for real-time detection of human motion relay on 3D CNN and ResNet 18. And the developed network is pre-trained on kinetics 400 datasets. The outcome of this work has been concluded that 3D CNN with ResNet-18 can improve the reliability of the network and decrease the losses in training and validation phases. Also, it can be concluded that the use of kinetics 400 datasets avoids the overfitting issue in the network.

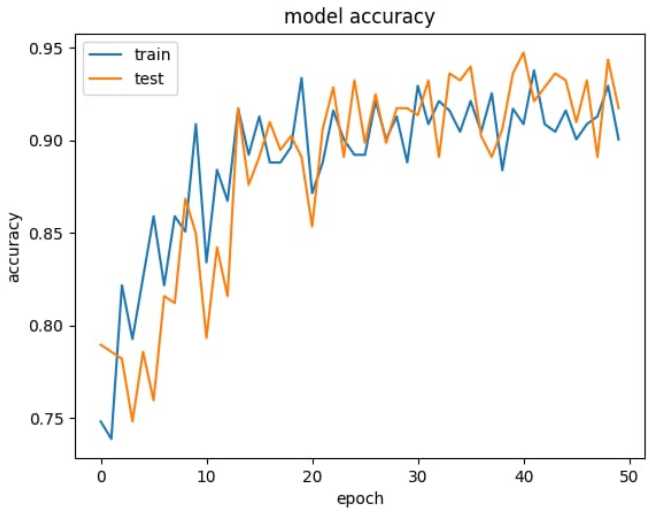

Fig. 7. ResNet-18 Model accuracy.

REFERENCES

[1] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," in IEEE Access, vol. 6, pp. 17913-17922, 2018, doi: 10.1109/ACCESS.2018.2817253.

[2] Q. Li, W. Yang, X. Chen, T. Yuan and Y. Wang, "Temporal segment connection network for action recognition," in IEEE Access, vol. 8, pp. 179118-179127, 2020, doi: 10.1109/ACCESS.2020.3027386.
[3] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition", In ECCV, 2010.
[4] S. Sharma, R. Kiros, and R. Salakhutdinov. (2015). "Action recognition using visual attention." [Online]. Available: https://arxiv.org/abs/1511.04119
[5] H. Wang, A. Kläser, C. Schmid and C. Liu, "Action recognition by dense trajectories," CVPR 2011, Colorado Springs, CO, USA, 2011, pp. 3169-3176, doi: 10.1109/CVPR.2011.5995407
[6] J. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),Jun. 2015, pp. 4694–4702.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
[8] Basly H., Ouarda W., Sayadi F.E., Ouni B., Alimi A.M. (2020) CNN-SVM learning approach based human activity recognition. In: El Moataz A., Mammass D., Mansouri A., Nouboud F. (eds) Image and Signal Processing. ICISP 2020. Lecture Notes in Computer Science, vol 12119. Springer, Cham. https://doi.org/10.1007/978-3-030-51935-329
[9] H. Kuehne, R. Stiefelhagen, T. Serre, and H. Jhuang, "HMDB51: A large video database for human motion recognition," in High Performance Computing in Science and Engineering. Berlin, Germany: Springer,2013, pp. 571–582.
[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
[11] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional neural networks for human action recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.
[12] X. Ou et al., "Moving object detection method via ResNet-18 with encoder–decoder structure in complex scenes," in IEEE Access, vol. 7, pp. 108152-108160, 2019, doi: 10.1109/ACCESS.2019.2931922.
[13] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. arXiv preprint, arXiv:1507.02159, 2015.
[14] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics human action video dataset. arXiv preprint, arXiv:1705.06950, 2017. 2, 3, 5, 6, 7