International Journal of Research in Engineering, Science and Management
Volume-2, Issue-5, May-2019
www.ijresm.com | ISSN (Online): 2581-5792

Human Activity Recognition using Convolutional 3D Network

C. Anjali[1], M. V. Beena[2]
[1]*M.Tech. Student, Department of Computer Science and Engineering, Vidya Academy of Science and Technology, Thrissur, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Vidya Academy of Science and Technology, Thrissur, India*

***Abstract*: Activity Recognition (AR) also referred as Action Recognition, is an elementary problem in emerging field of Computer Vision (CV). Specifically, Human action Recognition (HAR) is a problem in CV, based on classification and it involves the process of identifying the human movements. Nowadays it is a growing area of research and has gained popularity because of its wide range of applications including robotics and video surveillance. With the blooming of science and technology, the scientific milestones in the field of HAR are been achieved more rapidly. Today, most of the recent successful studies in this area of HAR are focused on Deep Learning (DL). In the field of Image recognition, DL techniques have outperformed the other techniques. Thus, this work aims to develop a 3D CNN for the Human Activity Recognition from videos. The model extracts features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the information encoded in multiple adjacent frames of the video. Namely three activities - Boxing, Hand waving and Running, from KTH Dataset is used for the purpose of HAR.**

***Keywords*: Human Activity Recognition, 3D network**

1. Introduction

A time when a robotic assistance will be able to understand as well as recognize the human activities, in a way that it can get the things done such that it can get things done without human intervention, is not very far enough. This can happen in near future, as our technology is developing at a rapid pace. In this review, we go through existing techniques for activity recognition that uses Deep Learning technique. Computer Vision (CV) comes under the field of AI, and it combines IP and ML and HAR is one of the basic problems in CV. Human activity recognition (HAR) can be defined as a classification problem used for the purpose of recognizing human movements.

Today, most of the recent successful studies in this area of Activity Recognition are mainly focused on Deep Learning techniques. We now witness a significant advancement in the problems being solved using the deep and data driven architectures. One of the most successful approaches for feature learning in complex data is accomplished using Deep Learning. It can discover the effective and valuable patterns in complex and large datasets. The Deep neural networks, like CNN have become the widely used method for the purpose of image content learning. Most recently, the advanced techniques that rely on CNNs, produced impressive improvement in the results compared to the techniques based on traditional features.

A. Why Activity Recognition?

As of late, HAR has turned into an intriguing issue since it has an extensive variety of utilizations. In the Engineering standpoint, AR has a wide scope of uses in the general public. The fields of applications incorporates Elderly supervision, Human-Computer association, Video reconnaissance, Medical conclusion, Learning for applied autonomy, Sports examination, Web video pursuit and its recovery [8].

B. Why Deep Learning?

DL is one of the fruitful methodologies for highlight learning in complex information. Deep neural systems like CNN have turned into the strategy for decision in learning contents in image. As of late, CNNs have been broadly utilized effectively to take care of HAR issues with regards to identification of gestures and exercises of day by day life. In examination with the standard methodologies, CNNs join advantageously the feature extraction and classification in a conclusion to-end approach. The feature extractors are non-straight changes, and they are gained straightforwardly from raw information being more discerning regarding human action classes. Interestingly, significant carefully assembled

highlights are difficult to figure and to scale. By stacking a few channels and pooling tasks, CNNs remove progressively

fundamental and complex human developments, taking in their non-straight and temporal relations. These structures share for all intents and purpose convolution and pooling tasks that are completed along the time pivot.

2. Related works

In Literature, several works are discussed that propose the recognition of human activities. The first approaches used to integrate human activities were based on the human joints trajectories (Campbell and Bobick, 1995; Niyogi and Adelson, 1994). These methods need specific techniques to detect parts

of the body or to track them in each image. Another approach that has already been explored is based on Bag-of-Words (BoW) [2], this type of proposal requires a large storage space for the less frequent features, besides the need to combine with other techniques for classification and extraction of Features.

With the success of Deep Networks, in special the AlexNet in 2012 [3], the deep models are being explored recently by different researchers. Among the deep architecture models, Convolutional Neural Networks (CNNs) gained attention because of their ability to learn contextual relationships between features [3]. This type of architecture has already achieved great results in domains such as digit recognition, speech recognition, emotion recognition.

Some authors have tried to apply the CNN for action recognition in videos. In the work of Ji et al. [4], they proposed the use of Convolutional Neural Network with two flows for the recognition of the actions, the first is the raw frame and the second is the optical flow with the temporal information of the movement between the frames. In the work of Wang et al. [5], they have developed a two-channel Convolutional network based on RGB frames. The first input of the model is the raw image and the second is the optical flow extracted from the motion in order to predict the trajectories. In the work of Lin et al. [6], they use a CNN to decompose temporal information by separating groups by sub-actions into RGB-D videos.

In the work of Chron et al. [7], in order to determine the action in videos, they proposed a descriptor based on body posture. Initially, they calculated the optical flow between frames and motion in each x and y direction. With the information of the directions, they calculate the movement of each part of the body. Finally, they use the raw frame and the flow of motion as input to a CNN for action recognition. Other authors have analyzed the option of using the 3D Convolutional Neural Network (3DCNN) for the recognition of human activities in videos. This kind of architecture can process temporal information, what has meaning for applications in videos [8]. In the work of Ji et al. [9], they used gray scale images, gradient and the optical flow along the x and y axes, taking these values as input in a 3DCNN for the recognition of actions in secure videos.

3. Architectural design

This session discuss about the proposed architecture and the convolutional model used for the classification of the video input.

A. Architecture

The proposed model uses 3D Convolutional Neural Network for the purpose of video classification. Figure 1 is an overview of the model. Currently, there are several proposals of Neural Networks for video classification: Ji et. al [11], Wang et. al [9], [10]. However, among these collected works, they all used some additional artifice to capture the temporal information. Next, we begin the discussion of our model. The model can be split into two parts. The first part is used to select the model settings. We select a dataset and apply the pre-processing steps to fit the data for the model.

Then we analyze the influence of the input on the model and its stability. Secondly, we load a random video input into the classification process and check its correctness.

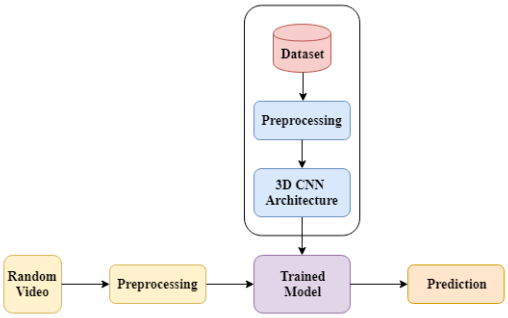

Fig. 1. Architecture of the system

B. 2D Convolutional Neural Networks

The Convolutional Neural Network (CNN) is biologically inspired by the Multilayer Perceptron [6] and in studies of the Hubel and Wiesel (1968) [14] with the visual cortex of cats. Hubel and Wiesel found that visual cortex contains complex arrangement of light-sensitive cells called receptive fields. These cells when act a local filters on input extract patterns that have strong local spatial correlation in the images.

Based on this, the CNNs explore the space-local correlation between the local patterns with the neurons of the adjacent layers. Each layer inert to variations outside of its open field as for the retina. Along the same way, the architecture guarantees that the filters learned produce the most grounded reaction to a spatially local input design. Each filter of the previous layer is represented as visual field for the next layer, this replication of the values share the same parametrization for the feature maps.

A CNN consists of at least one Convolutional layers (frequently a Subsampling layer) and after that by at least one full connected layer, in the standard version of a Multilayer Neural Network as proposed by LeCun te al. [6]. By form the values of each unit at position (x, y) of jth on the map of features of the layer ith, denoted by vxyij , is given by:

$$v_{ij}^{xy} = tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right) \quad (1)$$

where tanh(.) is hyperbolic tangent function, bij is bias for the features map with m indexes on the feature map set of layer (i-1)th connected to the current layer, wij pq and h the value of (p, q) of the connected kernel of the features map, and Pi and Qi are respectively, the height and width of the kernel applied.

In the sub-sampling layer the feature map is reduced by a pooling over the local neighborhood of the feature map to the next layer, the pooling operation can be performed with area overlay or not. CNNs can be constructed by multiple layers of

convolution and sub-sampling alternately.

C. 3D Convolutional Neural Networks

3D CNN can calculate the features for both temporal as well as spatial dimensions. 3D convolution is obtained through performing the convolutions of a 3D kernel on the cube formed by several frames continuously together. For this construction, the features map in the convolution layer needs to be connected with multiple frames of the next layer, thus capturing the motion information. The Figure 2, clearly shows the differences between the 2D and 3D models.

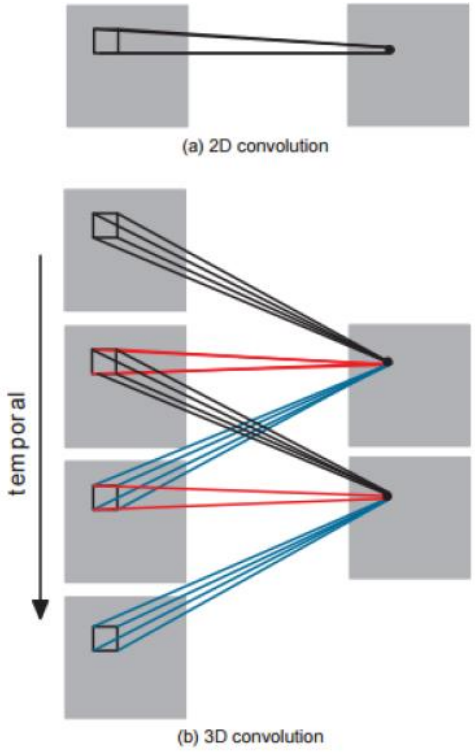

Fig. 2. Convolution

Fig. 2. We compare the 2D (a) and 3D (b) convolutions models. As can be noted in (a) have only spatial information represented by only one block, as in (b) the time kernel is the third dimension. The set of connections in red are share weights. In 3D CNN some kernel is applied with the overlay creating a 3D cube.

Formally the values of the positions (x, y, z) in jth of the map of features for each layer ith given by equation 2, which is the size of the 3D kernel along the temporal dimension, wijm pqz is the (p, q, r)th value of kernel connected in mth of the features map of the previous layer.

$$v_{ij}^{xyz} = tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \quad (2)$$

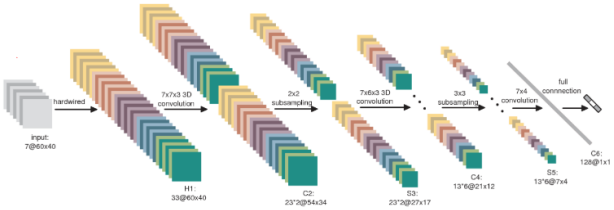

Fig. 3. A 3D CNN Architecture

D. Network Implementation

The architecture used is illustrated in Figure 3. The neural network was implemented using Keras. It consists of four Convolutional layers and four Max-pooling layers arranged alternately. We apply the kernel 3x3x3 to all the convolution layers and for the Pooling layers, we use a 2x2x2 window. It is followed by a Global Average pooling layer, then a full connected layer with 32 output neurons followed by hidden layers Dropout of 0.5. Finally, another Full connected layer as the output layer, with 3 output neurons.

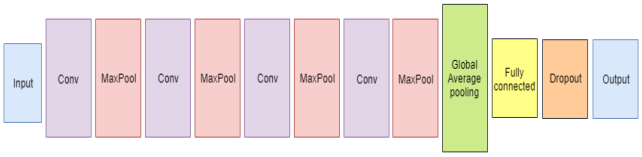

Fig. 4. Proposed Model

4. Experimental evaluation

In this section, we first describe the hardware and software settings used in the runtime environment. Secondly, we talk about the database used in the paper. We explain the process used to select the model settings and finally, we describe the approaches followed for the recognition of human activities in videos.

All experiments were performed on a desktop computer with the Windows 10 Operational System, processor core i3 and Colaboratory, which is a free jupyter notebook environment. All codes used were in the Python language version 3.5 in the Keras framework for Deep Learning. For using the whole dataset, ie, the 6 activities, the requirements include processor core i7 4790k with 4GHz frequency, Nvidia 970 GTX video card with 4GB of video and 32GB of RAM.

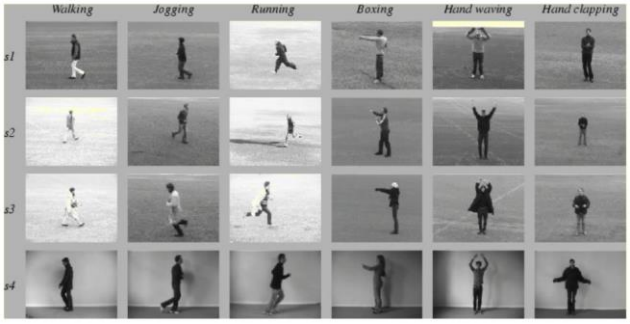

Fig. 5. Dataset

A. *Dataset*

The KTH dataset is used for recognizing human activities. The video dataset contains six types of human actions (boxing, handclapping, handwaving, jogging, running and walking) performed several times by 25 different subjects in 4 different scenarios. The videos were captured at a frame rate of 25fps and each frame was down-sampled to the resolution of 160x120 pixels. The dataset contains 599 videos 100 videos for each of the 6 categories (with the exception of Handclapping having 99 videos).

The proposed system implements Human activity recognition for three activities, namely - Boxing, Handwaving and Running.

B. *Materials and Methods*

The aim of this project is to create a model that can identify the basic human actions like running, jogging, walking, clapping, hand-waving and boxing. The model will be given a set of videos where in each video, a person will be performing an action. The label of a video will be the action that is being performed in that particular video. The model will have to learn this relationship, and then it should be able to predict the label of an input (video) that it has never seen. Technically, the model would have to learn to differentiate between various human actions, given some examples of these actions.

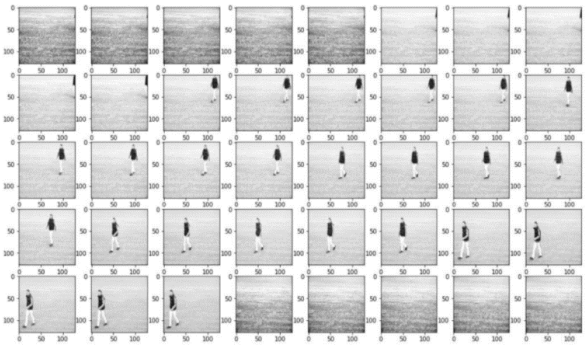

Fig. 6. The frames of a sample video of Walking

Figure 6 shows that in the videos of activities involving the movement of the whole body, a lot of frames in the video might be empty (no person performing any action). Also, if the person is moving very slowly, then most of the frames would be redundant. It would be a major challenge for the model to deal with such a problem.

The Data pre-processing stage consists of reading in the video frame-by-frame. The videos were captured at a frame rate of 25fps. This means that for each second of the video, there will be 25 frames. We know that within a second, a human body does not perform very significant movement. This implies that most of the frames (per second) in our video will be redundant. Therefore, only a subset of all the frames in a video needs to be extracted. This will also reduce the size of the input data which will in turn help the model train faster and can also prevent over-fitting.

5. Results

The model was trained on the three activities, namely - Boxing, Handwaving and Running, for 50 epochs. The weights of the model which gave the best performance on the validation data were loaded. The model used NADAM as the optimizer (instead of ADAM). In Keras, the default values of learning rate for ADAM optimizer is set to 0.001.

For NADAM, the default value of learning rate is 0.002 and there is a scheduled decay of learning rate. Given below is the learning curve of the model over 50 epochs.

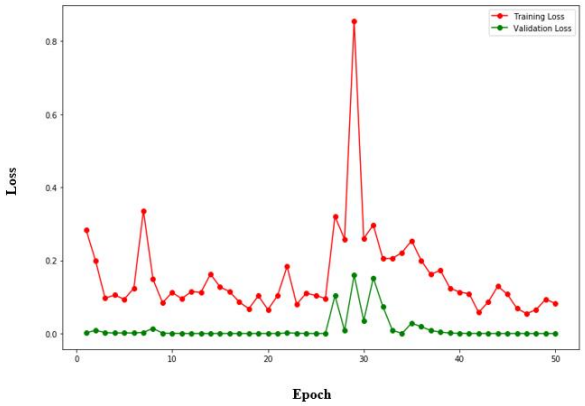

Fig. 7. Loss curve

Once the model has been trained on the training data, its performance will be evaluated using the test data. The model was then tested on the test data and it gave an accuracy of 98.48% on the test data.

	Precision	**Recall**	**f1-score**	**\Support**
0	1.00	1.00	1.00	31
1	1.00	1.00	1.00	32
2	1.00	1.00	1.00	29

Fig. 8. Performance measures

We have a Confusion Matrix in Figure 8. The labels on the vertical axis indicate the true class labels and the horizontal axis indicates the ordered classes. The main diagonal of the matrix represents the cases of predicted correctness examples, and the values outside the diagonal represent the values that the model predicts in the wrong way.

For the purpose of prediction, a random video for any of the three activities, is given as input to the trained model. On processing the input video, the corresponding class label is obtained as the output.

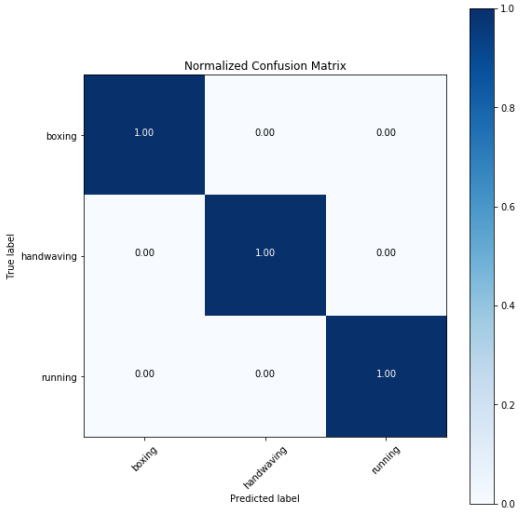

Fig. 9. Confusion Matrix

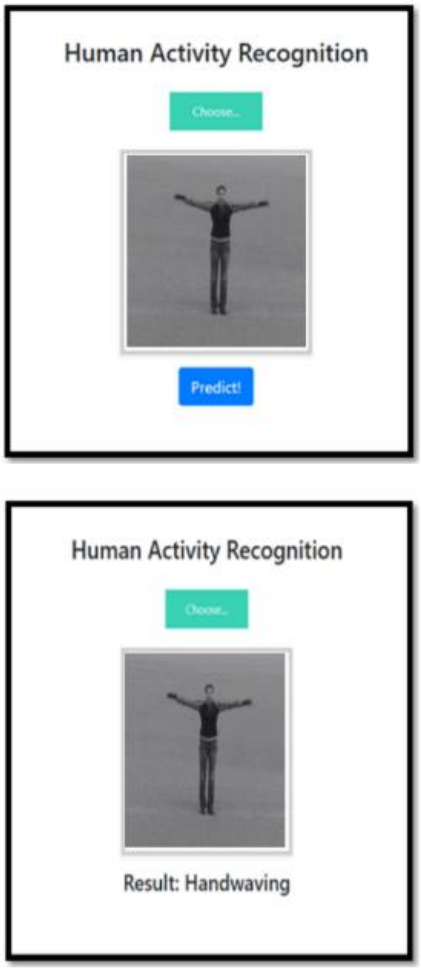

Fig. 10. Prediction Result

6. Conclusion

HAR can be considered as the ruler of video analysis issues because of its wide applications and the intricacy of the motion patterns created by verbalized body developments. The aftereffect of different research thinks about shows that the achievement of AR issues exceedingly relies upon a proper feature extraction process. In this article, we analyze the performance by using 3DCNN model for the recognition of human actions. The results obtained show the effectiveness of this type of solution in the classification of videos.

The article aims in developing a system for Human Activity Recognition from videos. The model uses the 3D CNN, which extracts features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the information encoded in multiple adjacent frames of the video. Three classes of activities - Boxing, Handwaving and Running, from KTH video dataset was used for training and testing the system. The results obtained show the effectiveness of this type of solution in the classification of videos. The system achieved test accuracy of 98.48 percent and works well for predictions. Thus, found that 3DCNN models have high capacity to process space temporal information.

References

[1] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3D Convolutional Neural Networks for Human Action Recognition", IEEE 2017.
[2] P. Foggia, G. Percannella, A. Saggese, and M. Vento, Recognizing human actions by a bag of visual words in Systems, Man, and Cybernetics
[3] (SMC), IEEE 2013.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Image net classification with deep convolutional neural networks 2012, pp. 1097-1105.
[5] Karen Simonyan, Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", Advances in neural information processing systems, 2014.
[6] L. Wang, Y. Qiao, and X. Tang, Action recognition with trajectory pooled deep-convolutional descriptors," IEEE 2015.
[7] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, A deep structured model with radiusmargin bound for 3D human activity recognition International Journal of Computer Vision, 2016.
[8] G. Ch eron, I. Laptev, and C. Schmid, P-cnn: Pose-based cnn features for action recognition IEEE 2015.
[9] S. Ji, W. Xu, M. Yang, and K. Yu, ,3d convolutional neural networks for human action recognition IEEE 2013.
[10] S. Ji, W. Xu, M. Yang, and K. Yu, 3d convolutional neural networks for human action recognition IEEE 2013.
[11] Wang, Y. Qiao, and X. Tang, Action recognition with trajectory pooled deep-convolutional descriptors 2015.
[12] G. Cheron, I. Laptev, and C. Schmid, P-cnn: Pose-based cnn features for action recognition IEEE 2015.
[13] S. Ji, W. Xu, M. Yang, and K. Yu, 3d convolutional neural networks for human action recognition IEEE 2013.
[14] Joe Yue-Hei Ng, Matthew Hausknecht, and Sudheendra Vijayanarasimhan, "Beyond Short Snippets: Deep Networks for Video Classification", IEEE 2015.