Deep Learning based Human Action Recognition

Ritik Pandey[1],*Yadnesh Chikhale*[2],*Ritik Verma*[3] and *Deepali Patil*[4]

[1,2,3,4]Ramrao Adik Institute of Technology, Information Technology Department, 40076 Nerul, India

Abstract. Human action recognition has become an important research area in the fields of computer vision, image processing, and human-machine or human-object interaction due to its large number of real time applications. Action recognition is the identification of different actions from video clips (an arrangement of 2D frames) where the action may be performed in the video. This is a general construction of image classification tasks to multiple frames and then collecting the predictions from each frame. Different approaches are proposed in literature to improve the accuracy in recognition. In this paper we proposed a deep learning based model for Recognition and the main focus is on the CNN model for image classification. The action videos are converted into frames and pre-processed before sending to our model for recognizing different actions accurately..

1 Introduction

The recognition of human activity is a primary issue in the field of computer vision. A human activity can be as simple as throwing a ball or brushing your teeth. At present, researchers have been working on this issue since it has received sensible attention. This work has its focus on recognizing an individual action. Deep Learning based human action recognition has been proposed in this work. The existing approaches are computationally more expensive so that makes the architecture more computationally efficient. Many different techniques can be used to recognize human action and research on this topic increased since the rise of popularity in artificial intelligence and machine learning. The basics of human action recognition are extracting features and prediction of action from an image or video.

CNN(Convolutional Neural Networks) is a particular deep learning method used for extracting and learning about features from Images. Currently there is no particular method for Video Classification. In this proposed system, the use of a CNN model to predict action which works by learning of features from individual images. It is difficult to perform human action recognition on a still image because of the limited resource of information, so using a collection of images from every video is proposed in our system where Video can be converted into a sequence of images. This work can be used in scenarios where a human activity is to be recognized by a computer.[3] Activity recognition is used in applications such as surveillance, anti-terrorists and anti-crime securities as well as life logging and assistance which reduces the cost of human resources. It is difficult to recognize human activity by a machine. This is an important challenge to the field of computer vision. The implementation of proper data pre-processing techniques has a high effect on the learning process of CNN model. These properties make the proposed method more suitable for action recognition in videos.

2 Related Work

Du Tran et al. have trained their data on large video datasets. They have used UCF101 Dataset for training the data from which they have achieved 52.8% accuracy with 10 classes [1].

Fernando Moya Rueda evaluated a novel CNN architecture for HAR (Human Action Recognition) using multichannel time-series acquired from body-worn sensors- IMUs. CNN-IMU [3]. This method is costly compared to normal non-sensors based recognition.

Most of them have emphasized their work mostly on human action recognition(HAR).In [2] they have to take a low-resolution life-logging camera data and predicted the action capture the camera, before that they have trained a large dataset Imagenet for the getting maximum percent of accuracy [20].

Y. Du, W. Wang, and L. Wang, et al presented a skeleton based action recognition for an end-to-end hierarchical recurrent neural network and they first split the human skeleton into five parts and then give them to five subnets. Only from the skeleton joins the alike human actions are not easily distinguished [4][5]. This method

* Corresponding Author: ritikpandey5@gmail.com,yadnesh1806@gmail.com ,vermaritk@gmail.com, deepali.patil@rait.ac.in

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

is computationally heavy and more complex.

Sometimes people get robbed or violence breaks down at a crowded place, at that moment it gets difficult to find the culprit or to keep an eye on the culprit. For keeping track of people at crowded places deep learning model are used for crowd management and for keeping track on suspicious activity[6].

According to the search, most of the research on action recognition is done on the state-of-the-arts [7] [14] and human action recognition (HAR) which is further used for the prediction of activity. The plotting of the graph is also done on the basis of the accuracy of the data [8] [9]. It explains the formation of the deep architecture with the help of Graph Estimation Procedure [10].

In depth based action recognition they have used 120 different classes for 3D based Human action recognition in theses they have evaluated the activity analysis [11]. It has datasets of 10s which were taken from youtube, in these there are various different action classes which have more interaction between humans with other objects. They have explained the statistics of dataset and performance for neural networks and they trained and tested the action classes dataset [13].

Gundong Guo et al. presented an overview of the state-of-the-art methods of still image-based action recognition and described and categorized on many high-level cues and low-level features for action based analysis for the still images. Still image-based human action recognition mainly focuses on recognizing a person's behaviour or action based on a single image [15]. In this work action recognition from still images is done and is not preferable for movement involving Human Action videos.

Various methods are used by researchers in the field of action recognition. These methods are of different types like Sensors based recognition [16], Machine Learning based action recognition and Deep Learning based action recognition.

Hong-Bo Zang & group has presented a review on the human action recognition method and also provided comprehensive overviews of this approach. They also included progress in hand-designed action features in RGB [17].

Fangyu Liu comes up with a 3DCNN-DQN-RNN method that combines all three methods to get the efficient semantic parsing of large-scale 3D point clouds. The main feature of their method, it provides the automatic process that plots the raw data to the classified results [18]. This method is a computationally expensive method and requires strong hardware to work.

Jun Liu, introduces a large set of data for RGB+D human action recognition. Their model helps in the long-term temporal correlation of the action for each part of the body and helps in providing them better action classification[19]. In this paper, skeleton data in the form of graphs is used which is a different form of data compared to video and images.

3 Methodology

To proceed with our proposed method of implementation, there are few basic steps. Fig 1 shows the flow of the steps to be carried out for the implementation of our proposed algorithm.

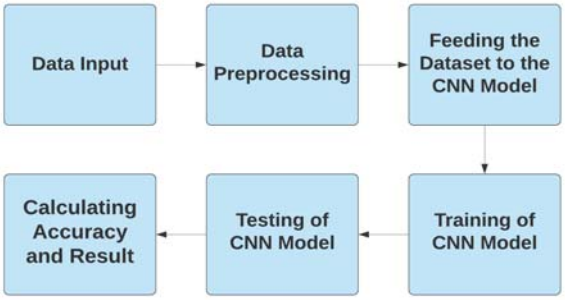


Fig.1. Flow Chart of Proposed System

A. Data Pre-processing: The first step is learning about the UCF101 dataset [1]. It is done by creating a data frame which has information about every video file in the dataset and which human activity class it belongs to. After that, the videos in the dataset are split into frames(images). The video which is used from the UCF101 dataset has 25 fps property. While converting into images, every 3rd frame for the 25 frames in 1 second is considered and saved. Later, the Image Data Generator by Keras rescales, shears, zooms and horizontally flips the frames so it'll be better for the learning process in the CNN model [12].

B. Training Model using Convolutional Neural Networks (CNN): Convolutional Neural Networks are used for learning and finding patterns in images. The CNN model is perfect for image classification. Each single image is a combination of pixels which are represented in the form of a matrix. The operations are applied on this matrix. A CNN has various layers namely, convolutional, pooling, fully connected, etc. In each of these layer's unique operations are applied in the process of learning. For image processing, our propose to use Tensor Flow and Keras which will be used for training the classifier. For other mathematical calculations and arrays, NumPy is used in Python. Pandas, matplotlib, and cv2 are used for

handling data frames, graphs and dealing with images respectively.

The structure of our CNN Model is:

1. Two sets of convolutional and max pooling layers.
2. Flatten layer.
3. Two dense layers.

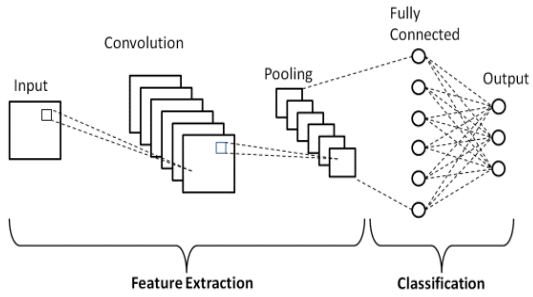


Fig.2. Sequence of layers in CNN Model

The image passes through multiple filters in two sets of convolutional 2D layers. The size of filters is kept constant and the number of filters is increased in the next convolutional 2D layer. In both these convolutional 2D layers, Relu or Rectified Linear Activation Function is used.To calculate the losses, Categorical_Crossentropy is used. Adam optimiser is used on the data to improve learning and increase accuracy.

C. Predicting Results: Accuracy and loss factors are calculated in the evaluation of the Train and Test dataset to see how efficient the model is. Training accuracy and loss graphs are made using matplotlib to know the structure of learning by epochs. The softmax probability output from the last dense layer is used to create a prediction probability matrix. Accuracy of every class is calculated separately. Random images of a few action classes are used from Google to test on the model. The model gives 80-88% accuracy currently.

4 Experimental Results

The paper gave information about the importance of Data Pre-processing. The dataset chosen for this is UCF101. This dataset has 13,000 videos which are divided into 101 categories or action classes. In total, the dataset is 27 hours. Each video has a length of 4 to 11 seconds. The variety of data available in the UCF101 dataset helps the model on testing data which are from outside of the dataset. The action classes in the dataset are Cricket Bowling, High Jump, Long Jump, Playing Piano, Yoyo, Playing Guitar, Ice Dancing, Punch, Drumming, Horse Race, Pull Ups, Volley Ball, Base Ball, Sumo Wrestling, Diving, etc. Noise data reduces the efficiency of the model and cleaning it can improve the accuracy. Using the techniques mentioned an accuracy of 80-88% on our CNN model is achieved. Different combinations of 6000 to 12000 images in datasets have been used in training and testing by CNN model. Various Video to Images conversion patterns are followed in pre-processing and the best learning and highest accuracy was found when the number of images of a video per second was 20-40% of the frame rate (eg. For a 30fps video, each frame that is a multiple of 3 is chosen to be saved as an image).

It was seen that as the number of epochs increased, the training accuracy increased and training loss decreased [10]. They both stop increasing and decreasing after a certain number of epochs.

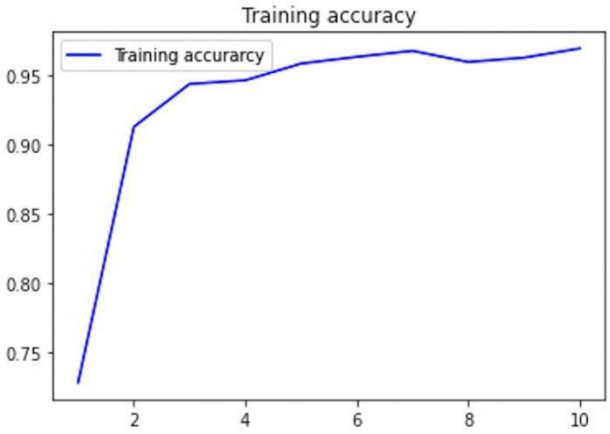


Fig.3. Training accuracy graph for 8000 images train set with 10 epochs

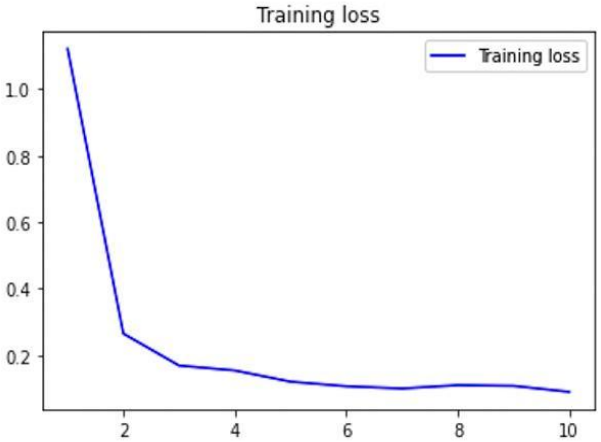


Fig.4. Training loss graph for 8000 images train set with 10 epochs

After the training of CNN model on various combinations of train datasets from 4000 to 9000 images, accuracies ranging from 80-88% were found on Evaluation.

Later, accuracies of all action classes were calculated and it was found that classes which had lower resolution and quality of video had lower accuracy. It was also found that Action Classes which has some kind of same and small similar activity in between, had an slight error in differentiating classes (eg. Athletic events like Pole Vault, Long Jump and High Jump have the same first half i.e. Running, in the similar background environment

```
model.fit(train_generator, epochs= 5)

Epoch 1/5
109/109 [==============================] - 831s 7s/step - loss: 2.3894 - accuracy: 0.5453
Epoch 2/5
109/109 [==============================] - 55s 501ms/step - loss: 0.3157 - accuracy: 0.8967
Epoch 3/5
109/109 [==============================] - 55s 501ms/step - loss: 0.2059 - accuracy: 0.9327
Epoch 4/5
109/109 [==============================] - 56s 510ms/step - loss: 0.1580 - accuracy: 0.9506
Epoch 5/5
109/109 [==============================] - 55s 506ms/step - loss: 0.1503 - accuracy: 0.9489
<tensorflow.python.keras.callbacks.History at 0x7fd26a8ed450>

[ ] model.evaluate(test_generator)

36/36 [==============================] - 237s 7s/step - loss: 0.3927 - accuracy: 0.8837
[0.3926926553249359, 0.883700430393219]
```

Fig.5. Accuracy of 88.37% on Evaluation of test dataset

To be sure about the efficiency of the model, random images of the respective class from Google were given for prediction to our CNN model. These random images were predicted correctly as shown in Figure 6, 7 and 8.

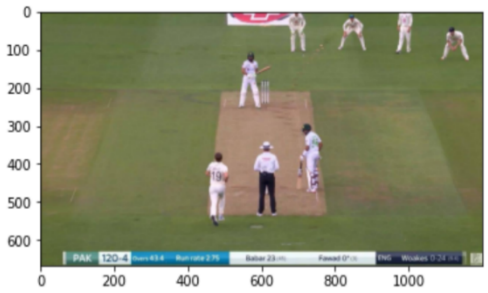
```
img = cv2.imread(os.path.join("/MyDrive/",'pool.jpg'))
imgplot = plt.imshow(cv2.cvtColor(img, cv2.COLOR_BGR2RGB))
plt.show()
```


```
print(predictionz)

['Diving']
```

Fig.6. Class Prediction on a random Diving image from Google

```
img = cv2.imread(os.path.join("/MyDrive/",'17-1.jpg'))
imgplot = plt.imshow(cv2.cvtColor(img, cv2.COLOR_BGR2RGB))
plt.show()
```




```
print(predictionz)

['CricketBowling']
```

Fig.7. Class Prediction on a random CricketBowling image from Google

```
img = cv2.imread(os.path.join("/MyDrive/",'850.jpg'))
imgplot = plt.imshow(cv2.cvtColor(img, cv2.COLOR_BGR2RGB))
plt.show()
```


```
print(predictionz)
```

```
['SumoWrestling']
```

Fig.8. Class Prediction on a random Sumo Wrestling image from Google

5 Future Scope

In the proposed system, the videos are converted into frames and are fed to the CNN Model. CNN is a deep learning neural network which can learn and find patterns in images. It has been seen that the videos were treated as separate images. Further works can be on adopting an approach that treats the video data more like a video and not like an image, eg. A combination of CNN and RNN. In the future for more accurate prediction, mapping of the Human Body in the form of graphs can be done using computer vision algorithms. Skeleton data in the form of graphs can be used. There are also different sensors which are used to create data which can help in better accuracy. 3D CNN is a different field which is also growing at present times. Study and work in these interesting Video Classification fields will help to make accurate predictions.

6 Conclusion

In this paper, Human Action Recognition Algorithm by using frames from Human Action videos is proposed. First, the Data Preprocessing takes place where videos are separated into frames. Then various operations are done on the images to make it ready for the CNN model. Features are extracted from the video frames, which are then used to help determine the accuracy of the model, evaluation, and prediction of classes of videos. Here, the output for small images are combined as a final output of the class. The variety of data in action classes of the UCF101 dataset helps in better prediction of images outside of the dataset. Since the whole image which includes both human character and background in the frames are used in CNN model, essential information is extracted. Because of the huge variety of data in the used UCF101 dataset, the proposed model can be tested on random videos which are separate from UCF101.

References

1. Du Tran, Lubomir Bourdev, Rob Fergus,Lorenzo Torresani, Manohar Paluri “Learning Spatiotemporal Features with 3D Convolutional Networks”, IEEE International Conference on Computer Vision (ICCV), 2015.

2. Tsai, Jen-Kai & Hsu, Chen-Chien & Huang, Shao-Kang. “ Deep Learning-Based Real-Time Multiple-Person Action Recognition System”, Sensors. 20. 4758. 10.3390/s20174758, Aug 2020.

3. Romaissa, Beddiar & Nini, Brahim & Sabokrou, Mohammad & Hadid, Abdenour, “Vision-based human activity recognition: a survey”, Multimedia Tools and Applications. 79. 10.1007/s11042-020-09004-3, Aug 2020.

4. Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition”, In IEEE Conference Paper on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1110– 1118.

5. S. Song, C. Lan, J. Xing, W. Zeng, “An End-to-End SpatioTemporal Attention Model for Human Action Recognition from Skeleton Data” in AAAI, pp. 4263–4270, 2017.

6. Riddhi Sonkar, Sadhana Rathod, Renuka Jadhav Deepali Patil et al. "Crowd Abnormal Behaviour Detection using Deep Learning", ITM Web of Conferences, 2020

7. T. S. Kim and A. Reiter, “Interpretable 3d human action analysis with temporal convolutional networks,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 10.1109/CVPRW.2017.207

8. M. Niepert, M. Ahmed, Konstantin Kutzkov et al. “Learning convolutional neural networks for graph,” in International Conference on Machine Learning (ICML), 2016.

9. J. Bruna, W. Zaremba, Arthur Szlam, Yann LeCun et al. “Spectral Networks and Locally Connected Networks on Graphs,” in International Conference on Learning Representations, 2014.

10. M. Henaff, J. Bruna, Yann LeCun et al. “Deep convolutional networks on graph-structured data,” arXiv:1506.05163 [cs.LG], 2015

11. Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, Alex C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019. 10.1109/TPAMI.2019.2916873

12. F. Caba Heilbron, B. Ghanem, J. C. Niebles, et al. .“A large-scale video benchmark for human activity understanding”. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 961-970

13. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, and others, “The Kinetics Human Action Video Dataset,” arXiv:1705.06950 cs. CV, May 2017.

14. Yu Cong ,Yun Fu “Human Action Recognition and prediction: A survey”, Computer Vision and Pattern Recognition, Cornell University, June 2018.

15. Gundong Guo, Alice Lai et al. “A survey on till image based human action recognition” west virginia university, may 2014.

16. Chen Chen, Roozbeh Jafari, Nasser Kehtarnavaz et al. “A survey of depth and initial sensor fusion for human action recognition” , 2017.

17. Hong Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du and Duan-Sheng Chen et al. “A Comprehensive Survey of Vision-Based Human Action Recognition Methods” feb 2019.

18. Fangyu Liu, Shuaipeng Li, Liqiang Zhang, Chenghu Zhou , Rongtian Ye , Yuebin Wang , Jiwen Lu et al. “3DCNN-DQN-RNN: A Deep Reinforcement Learning Framework for Semantic Parsing of Large-scale 3D Point Clouds” Tsinghua University.

19. Amir Shahroudy, Jun Liu, Tian-Tsong Ng and Gang Wang et al. “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis” Singapore Institute for Infocomm Research, april 2016.

20. Aksasse H., Aksasse B., Ouanan M.,”Deep Convolutional Neural Networks for Human Activity Classification”,. In: Jain L., Peng SL., Alhadidi B., Pal S. (eds) Intelligent Computing Paradigm and Cutting-edge Technologies. ICICCT 2019. Learning and Analytics in Intelligent Systems, vol 9. Springer, Cham, 2020.