



# Human Action Recognition Using VA-NN & 3D CNN

Jyoti D Biradar<sup>1</sup>, Rashmi S R<sup>2</sup>

Department of CSE, Dayananda Sagar College of Engineering, VTU, Bengaluru, Karnataka, India<sup>1,2</sup>

**ABSTRACT:** In Human Action recognition it is difficult to predict the movement of a person. In today's world, human action recognition has attracted attention of researchers. The one key challenge in human action recognition is there will be many dissimilarities in action representations when they are captured from different viewpoints. In order to make the effects of dissimilarities of actions, this introduces a scheme known as view adaptive neural networks and convolutional Neural Networks. which automatically determines the virtual observation viewpoints over the course of an action in a detailed manner.

Predicting the actions of human from videos and images. And feeding them to next level and displaying the required number of actions. Since human body is non-rigid, definition of body plane is 'hip', 'shoulder', 'neck', is not always suitable for orientation alignment. Hence rather than struggling to define complex criteria to handle non rigid body, the CNN model is used to automatically learn the shape and actions of human. Hence the actions can be identified using different viewpoints of a human from given videos and images.

**KEYWORDS:** Human action recognition, Human action prediction, 3D Convolutional neural network, Camera, 3D displays.

## I. INTRODUCTION

The main aim of HAR is to recognize the activities from many observations and the actions performed by the human in environment. It has many applications such as computer interaction, video indexing, human-computer interaction, game control, video games and video understanding. According to types of input data the human action recognition is categorized into RGB based and 3D -skeleton based approaches.[1] The key challenge in recognizing the action lies in the large variations of action representations when they are captured from different viewpoints. Computer vision is used to enabling machines to understand the human actions in videos and pictures.

One of the application of computer vision, human activity recognition includes is Action recognition : (recognized human action from video containing complete action execution).

Some authors have tried to apply the CNN for action recognition in videos. In the work of Ji et al. [2], 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.

There are many issues in human action recognition from videos such as:

- The angle of camera through which it captures videos is the main issue in human action recognition. In real world, the angle of camera keeps on changing and hence the performance of human action recognition should be invariant to change in angle of camera motion.
- Change in illumination and occlusion also affect the human action recognition.
- Due to less intraclass variation and high interclass variation create problems in different human action recognition.
- Appearance of human changes due to the way of performing actions keeps on changing based on the surface on which action is performed, clothes also play an important role in the appearance of human and the objects they carry with them. Therefore it is been a research issue to recognize human action irrelevant to the appearance of human[3].

Human action recognition is a step to understand and percept the nature which is a part of machine perception. The prediction is higher layer than the recognition in human actions, which give the machine the ability of imagination and reasoning[4].

This work focused on the human action classification and recognition. There are several approaches using the various dataset to solve action recognition problem with the certain hypothesis, which does not give a good practical value for the real-time environment. Initially, the specific measure is used to represent the features from the extracted raw video frames in order to identify the human actions. Where these assumptions will fail in giving the information in ending and it is difficult to identify the specific feature in actual scenario. The above problem can be solved partially by Deep



Learning models using approaches by feeding the input training data, which can learn the features and compute the results from the particular datasets without making any assumptions. Hence this paper focuses on identifying human actions more accurately using convolutional neural networks. CNN is a deep model that obtains complicated hierarchical features via convolutional operation alternating with sub-sampling operation on the raw input videos. Hence it ensures that CNN gains more accurate performance in visual target recognition tasks through a good adjustment during training. And CNN has invariance for a particular pose, illumination, and disorderly environmental change. The first attempt of using CNN was developing the 3D CNN model that extract features from both spatial and temporal dimensions by performing 3D convolutions, Hence the captured information is encoded in multiple frames so to generate multiple channel information from input frames. And final representation can be obtained by combining all channels information.

## II. RELATED WORK

In [5] authors motivate the need for and challenges involved in classification and recognition of temporal data resulting from Object tracking captured through different viewpoints. They cast view-invariant activity recognition problem as affine-invariant image shape retrieval. That approach detects the CSS contour maxima and represents the trajectories based on the locations of these peaks. The limitation of this approach is that it does not model the data between segmentation points. This short coming is alleviated in the CDF+PCA based representation scheme.

In [6] it is proposed a simple end-to-end but high-efficiency and high-precision framework for skeleton based action recognition. they represented human skeleton sequences as images to transform the temporal dynamics of sequences into the spatial structure information in images. Then a hierarchical architecture based on CNN was proposed for feature representation learning and classification.

By [7] , it is addressed that the importance of automatic understanding and characterization of human action. He proposed a 3D CNN model to recognize human actions. where model constructs features from both spatial and temporal dimensions by performing 3D convolutions. It has validated the approach on the KTH and JHMDB datasets. and shown that 3D CNN architecture can be a very useful tool for recognizing human action without the need for hand-tuned foreground segmentation or any preprocessing steps.

The problem of human action recognition by applying ConvNets to skeleton sequences. Proposed a method to describe the joint trajectories to three orthogonal JTMs to encode the spatial-temporal information into texture patterns. The three JTMs are complementary to each other. Such kind of image based representation enables to fine-tune the current ConvNets models are trained on image data for skeleton sequence classifications without training the deep ConvNets afresh[8].

## III. THEORY

### A. View Adaptive Neural Networks & Neural Networks

There are two view adaptive neural networks i.e. RNN and CNN, named VA-RNN and VA-CNN respectively. That is View adaptive recurrent neural network and view adaptive convolutional neural network. VA-RNN consists of view adaptation subnetwork and the LSTM for transforming the input data and recognizing the actions performed by the humans. Where as VA-CNN consists of CNN based view adaptation subnetwork and the main convolutional network (ConvNet). Each network is trained End-to End by optimizing the classification performance[1,9].

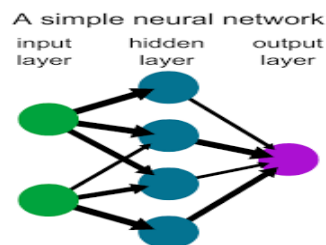


Fig. 1. A Simple Neural Network

A simple neural network consists of three layers such as 1) input layer 2) hidden layer 3) output layer .

Input layer consists of initial data for neural network. Hidden layer is the intermediate layer of the input and output layer where the computation is done. Output layer produces the results for given inputs[10].



B. Convolutional Neural Network

Convolutional neural network is a type of artificial neural network that is used in recognizing actions from images and videos and it is specifically designed for processing the pixel data[11]. A convolutional neural network can take in an input image and assigns the importance to aspects of image and then differentiate one from the other.

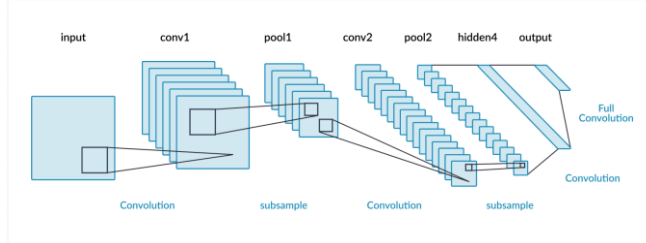


Fig. 2. Convolutional Neural Network

Figure shows a typical architecture of the convolutional neural network model. CNN is composed of several kinds of layers:

- Convolutional layer creates a feature map to predict the class probabilities for each feature by applying a filter that scans the whole image, few pixels at a time.
- Pooling layer nothing but down sampling which measures the amount of information generated by CNN for feature and maintains the most useful information.
- Fully connected input layer “Smoothens” the output generated by previous layer and then turns them into single vector so that can be used as input to next layer.
- Fully connected layer can apply weights over the input generated by feature analysis to predict an accurate label.
- Fully connected output layer generates the final probabilities to determine a class for the image.

IV. SYSTEM DESIGN

In order to train a model to recognize human actions, a dataset of videos is required. This would only require a binary classifier model that would predict the class probabilities for each feature by applying a filter that scans the whole image, few pixels at a time. A more interesting problem would be one where the human actions are captured from different views in different categories. like walking, standing, sitting, singing, playing and etc[12][13]. for example. This is a harder problem and requires a multi classifier model and more specific dataset.

The dataset consists of videos of humans performing many possible activities, which are shown in below figure while considering the dataset of UCF101[15].



Fig. 3. Different actions performed by human



The figure shows all the different type of categories of videos in the dataset[14]. This makes some categories easier to distinguish from others, while other predictions might be consistently labeled wrong due to their similarity to other categories.

A. Dataset

The dataset used in action recognition is UCF 101 data set which is realistic in environment and are collected from YouTube., having 101 action categories. This UCF101 data set is an extension of UCF50 data set which has 50 action categories, whereas UCF101 dataset has 101 categories with 13320 videos. The UCF101 dataset can gives many more imbalances in terms of actions[15][16] and the large variation can be in camera motion, object appearance, object scale, viewpoint, human pose, cluttered background, illumination conditions, etc. It is the most challenging data set till date. Most of the action recognition dataset are not optimistic or realistic. But UCF101 is aimed to give the realistic action recognition categories to the researchers for learning purpose.

The action videos in UCF101 action categories has 25 groups, where each group consists of 4-7 videos of an action. and the videos from same group may share common features, such as it can have same background, same viewpoint, etc.

The action categories can be divided into five types such as Human-object interaction, Body-Motion only, Human-Human Interaction, Playing Musical Instruments and Sports.[16]

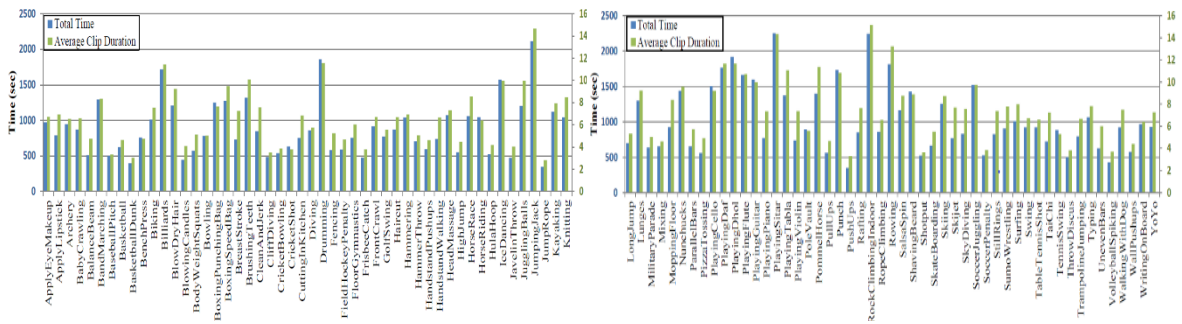


Fig. 4. For each class the Total time of videos is illustrated using the blue bars. The average length of the clips for each action is depicted in green colors.

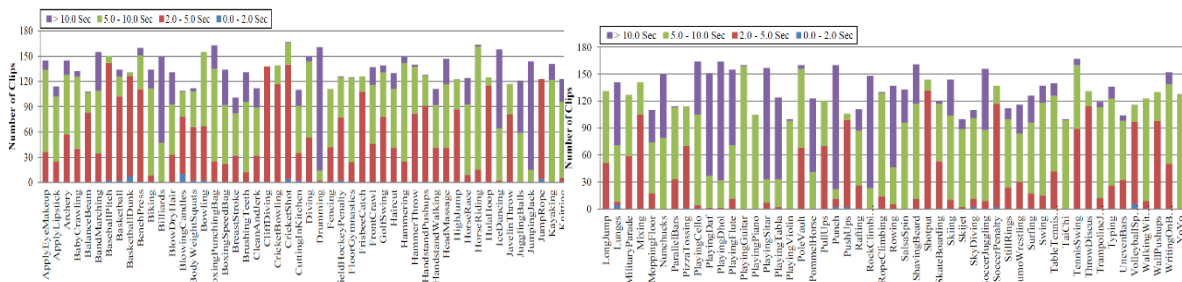


Fig.5 Number of clips per action class. The distribution of clip duration is illustrated by colors.

B. Feature Extraction

Feature extraction is done after the preprocessing phase. This phase should contain the information required to distinguish between classes, and it should be intensive to irrelevant changes in the input, and also be limited in number, to permit, efficient computation of discriminative functions and to limit the amount of data required for training. The data pre-processing method is used to process the dataset. “A dataset is a collection of data objects”. Data objects are created by a number of features. That express the basic attributes of an object. The features are either categorical or numeral. Feature selection is the method of selecting the features that has high rate of success and also removing the input variables that are not relevant to improve the performance and storage space of the system[17]. The features are selected by creating a connection between each input variable and the target variable using correlation stats and selecting the input variable that have the well built association with the target variable.



### C. Training and splitting

In this training dataset it contains the original dataset. the model is trained on the training dataset. The training dataset often consists of pairs of an input vector and the corresponding output vector. In this dataset[15] the video contains the different category videos which contains the human actions. The overall dataset contains close to 13320 videos that fall under the 5 categories and 25 groups shown in fig.3.. More specifically, the dataset was originally used is UCF101 which is extension to the dataset UCF50. The action videos in UCF101 action categories has 25 groups, where each group consists of 4-7 videos of an action. and the videos from same group may share common features, such as it can have same background, same viewpoint, etc.

### D. Validation

The validation generator can work as same as that of training generator. But the data of validation has no relation with data of training. Hence there is no need to separate validation batches according to training batches. Also, the total number of samples in training data is not related to the total number of samples in test data. The data sample is used to provide the unbiased evaluation of model on training dataset while tuning hyper parameters. Then evaluation becomes more biased on the validation dataset which is incorporated into the model configuration. Whenever the error on validation increases, the validation datasets can be used for the early stopping. The validation data set contains 20% of the training images to improve the accuracy in the model. The validation step is similar to steps per epoch but on the validation, data set instead on the training data. If you have the time to go through the validation data set.

### E. Prediction

Prediction refers to the trained algorithms a old dataset and applied to new data when predicting the possibility of a particular outcome. The data must be optimized and generalized which means that the data should give the best possible outcome and perform well on unknown data. The trained data predicts the output on the basis of the pattern of the datasets.

## V. IMPLEMENTATION & RESULTS

In the first attempt of predicting the human actions. The following snapshots defines the results or outputs that we will get after step by step execution of all the modules[17][18] of the system.

### Action Prediction:

Taking videos as model input and predicting the required number of actions based on their probabilities.

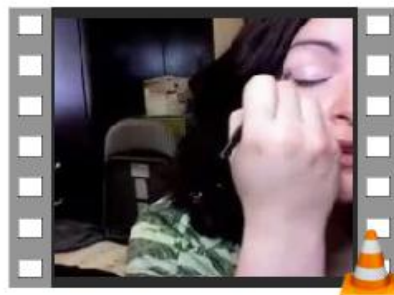
Function : `predict(sample_video)`

Using above function the number of actions are predicted and display using the `print()` function.

Following are some of the examples(snapshots) of action prediction from the given videos.

Using applyeyemakeup video predicting the actions such as:

```
video_path = fetch_ucf_video("v_applyeyemakeup_g01_c01.avi")
```



The above function fetches the video of applyeyemakeup from ucf dataset and predicts the top 5 actions using below function.



||Volume 8, Issue 6, June 2020||

```

▶ predict(sample_video)
↳ Top 5 actions:
  filling eyebrows      : 98.13%
  applying cream       : 1.57%
  waxing eyebrows      : 0.17%
  playing harmonica    : 0.07%
  brush painting       : 0.04%

```

### Loading and converting to gif from sample videos:

Predicting the top 10 actions of human activities and then converting the activity to the gif


```
video_path = "Formas_con_armas..ogv"
```

```
sample_video = load_video(video_path)[:100]
```

```
sample_video.shape
```

```
output: (100, 224, 224, 3)
```

```

[44] to_gif(sample_video)
↳ 

```

```

▶ predict(sample_video)
↳ Top 10 actions:
  side kick           : 58.31%
  dunking basketball : 10.59%
  high kick           : 4.24%
  playing basketball : 3.85%
  tai chi             : 3.49%
  shooting basketball : 3.48%
  playing badminton  : 2.52%
  playing cricket    : 1.76%
  playing tennis     : 1.40%
  cheerleading       : 1.27%

```

## VI. CONCLUSION AND FUTURE WORK

In Human Action recognition it is difficult to predict the movement of a person. In today's world, human action recognition has attracted attention of researchers. The one key challenge in human action recognition is there will be many dissimilarities in action representations when they are captured from different viewpoints. In order to make the effects of dissimilarities of actions, this introduces a scheme known as view adaptive neural networks and convolutional Neural Networks. which automatically determines the virtual observation viewpoints over the course of an action in a detailed manner.

In Human action recognition, recognizing the actions of human represents any number of actions from the given video. Predicting the video shape and number of actions from the sample video and display the required number of actions from the given dataset.

## REFERENCES

1. Pengfei Zhang ; Cuiling Lan ; Junliang Xing ; Wenjun Zeng ; Jianru Xue ; Nanning Zheng "view adaptive neural networks for high performance skeleton based human action recognition "in IEEE transaction on pattern analysis and machine intelligence, 2019
2. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems,



3. Chandni J. Dhamsania and Prof. Tushar V. Ratanpara “A Survey on Human Action Recognition from Videos,” in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*.
4. Meixia Fu, Na Chen, Zhongjie Huang, Kaili Ni, Yuhao Liu, Songlin Sun, Xiaomei Ma a human action recognition :survey,” in *2019 International Conference on Research in Intelligent and Computing in Engineering (RICE)*,
5. I. Bashir, A. A. Khokhar, and D. Schonfeld. View-invariant motion trajectory-based activity classification and recognition. *Multimedia Systems*, 12(1):45–54, 2006.
6. Y. Du, Y. Fu, and L. Wang. Skeleton based action recognition with convolutional neural network. In *ACPR. IEEE*, 2015.
7. Sameh Neili Boualia “3D CNN for Human Action Recognition” in 2018
8. Pichao Wang “Action Recognition Based on Joint Trajectory Maps with Convolutional Neural Networks,” in *2016*.
9. Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. *CVPR*, 2017
10. Rahmani and M. Bennamoun. Learning action recognition model from depth and skeleton videos. In *ICCV*, 2017.
11. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose.
12. Recognition in parts from single depth images. In *CVPR*, 2011 R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a lie group. In *CVPR*, 2014.
13. J. Wang, Z. Liu, and Y. Wu. Learning actionlet ensemble for 3d human action recognition. In *Human Action Recognition with Depth Cameras*, pages 11–40. 2014.