ISSN No:-2456-2165

Deep Neural Network Approaches for Video Based Human Activity Recognition

Chaitanya Yeole
School of Electronics and Communication
MIT World Peace University
Pune, India

Hricha Singh
School of Electronics and Communication
MIT World Peace University
Pune, India

Hemal Waykole
School of Electronics and Communication
MIT World Peace University
Pune, India

Anagha Deshpande
Assistant Professor
School of Electronics and Communication
MIT World Peace University
Pune, India

Abstract:- In this paper we explained, tried, and tested methods of Human Action recognition for the application of video surveillance. This paper provides a method for automatically recognizing human activities included in video sequences captured by a single large view camera in outdoor locations. The elaboration of the dataset which are videos are taken with the resolution of 720x480, precisely explained. The methods we implement are CNN-VGG16 model and the Single-frame CNN model. We demonstrated our techniques using real-world video data to automatically distinguish normal behaviors from suspicious ones in a playground setting, films of continuous performances of six different types of human-human interactions: handshakes, pointing, hugging, pushing, kicking, and punching. As per the observation, we concluded that the Single frame CNN model shows much better results as compared to CNN VGG16. The implementation was done in python. This paper consist of how the convolution neural networks' simple classification method proved to be efficient for the prediction of the activity by using a single frame method. The working of this method is briefly mentions in the methodology. The difference and the drawback of these methods for human activity recognition can be clearly seen in the output and results of the respective.

Keyword:- *Artificial Intelligence (AI) Models Are Created to Perceive the Movement of Human from the Provided Dataset.*

I. INTRODUCTION

Human activity recognition has a lot of importance in many applications including video surveillance, Human-human interactions, and human- computer interaction. Automatically recognizing activities of interest plays a vital part in many of the present video surveillance. Human activity recognition is one of the significant innovation to screen the dynamism of an individual and this can be accomplished with the help of Machine learning methods. This is can be used for security purposes and most importantly detect criminal activity within minutes. It has wide use of applications. Therefore, it is always desirable to develop new activity recognition algorithms and have the research of tried and tested methods by which we can get higher accuracy and stronger capability for handling various scenarios.

This study attempts to provide a comprehensive review of video-based human activity recognition, as well as an overview of various methodologies and their evolutions, by covering both typical classic works of literature and theories of possible solutions. One of the method used in activity recognition is CNN-LSTM, this method not only enhances the accuracy of predicting human actions from raw data, but it also decreases the model's complexity and eliminates the need for advanced feature engineering [1]. CNN is increasingly being used as a feature learning method for human activity recognition [2]. Extracting significant temporal features from raw data is critical. The majority of HAR techniques need a significant amount of feature engineering and data pre-processing, which necessitates domain expertise [3]. Many applications, like as video surveillance, health care, and human-computer interaction, are founded on vision-based HAR research (HCI) [4]. HAR faces numerous difficulties, like enormous fluctuation of a given activity, closeness between classes, time utilization, and the high extent of Null class. These difficulties have driven researchers to build techniques for methodical highlights and proficient acknowledgment strategies to viably take care of these issues [5]. Many writers have used sequential techniques and space-time volume approaches to express actions and perceive action sets directly from images. Then, for anomalous action states, they employed hierarchical recognition approaches. For hierarchical recognition, statistics-based approaches, syntactic methodologies, and description-based methodologies are all used [6]. Anomaly detection is one of the most well-known applications of human activity recognition [7].

II. DATASET AND PREPROCESSING

Hand-shaking, pointing, hugging, pushing, kicking, and punching are all examples of human-human interactions that can be seen in the videos. A total of 20 video sequences with a length of roughly one minute make up the dataset. Each video comprises at least one execution per interaction, resulting in an average of eight human activity executions per video. The videos have volunteers dressed in around 15 different ways. The videos were shot at a resolution of 720x480 pixels, with a person's height in the video being 200 pixels.

The videos are divided into two sets for pre-processing, which makes it easier to implement the dataset. Set 1 is made up of about ten video clips shot on a playground. Set 1's videos were shot at a different zoom rate. Set2, which consists of the remaining 10 sequences, was shot on a green lawn in a breezy environment. From sequences 1 to 4 and from 11 to 13, only 2 interacting volunteers with different clothing appear in each video. That video sequence has both conversing people and civilians from sequences 5 to 8 and 14 to 17. Sets9, 10, 18, 19, and 20 are pairs of interacting volunteers who participate in the activity at the same time. The background and scale of each set are distinct.

For the implementation of the dataset in themethods, we had to pre-process this data further. We converted the dataset into frames using various python libraries. Thesewere further segregated into train and test datasets for the application of the techniques i.e., VGG 16and CNN.

III. PROPOSED METHOD

1.1 VGG -16 (Visual Geometry Group)

VGG-16 is a large-scale image recognition architecture based on deep convolutional neural networks. The University of Oxford's K. Simonyan and A. Zisserman proposed this approach. The architecture of VGG16 has the input that is the pre- processed frames, to the network is an image of dimensions 224x224 x3. The first 2 layers have 64 compartments of 3x3 filter size and the same padding. After the max pool layer of dimensions 2x2, 2 layers have convolution layers of filter size 256 and 3x3.This is followed by a max-pooling layer of stride 2x2 which is the same as the previouslayer. This sequence is repeated twice.

Later 2 sets of 3 convolution layers and a max pool layer are formed. Each has 512 required for both the techniques was the same. The dataset was converted into frames and further divided into train and testcategories which were implemented on the models. More elaborated information aboutthe models:-filters of 3x3 size with the same padding. The frames are then sent to a two-layer convolution stack. The filters we utilize in these convolution and max-pooling layers are 3x3 in size. After the formation of the convolution and max- pooling layer, a 7x7x512 feature map was achieved. The output was flattened to makeit a 1x25088 feature vector. Three layers achieved were further passed to the Soft- max layer to normalize the classification vector. In this architecture, the activation function used is Relu as it is fast and more efficient and also reduce the probability of vanishing gradient problem. Considering the case of human activity recognition this model does not show the best results, it hasvarious speed limitations.

1.2 CNN Model

Convolution Neural Networks are a type of Neural Networks, which can identify and classify features from frames. Analyze visual images is one of the widely used functions. Video and picture identification, image classification, medical image analysis, computer vision, and natural language processing are just a few of the applications (NLP).

A CNN architecture is divided into two parts:-

- Feature Extraction is a process where a convolution tool identifies and separates the various features of the pre-processed frames for analysis.
- Based on the characteristics extracted in the previous phases, a fully connected layer recognizes the output from feature extraction and predicts the class of the frames.

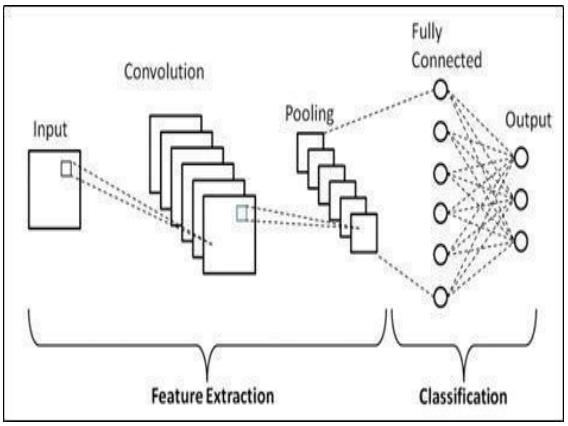

Fig.1 CNN Architecture

Convolution Layer: The first layer in the Feature Extraction. The mathematical equations of convolution are executed between the input frame and a particular size for a filter is also executed. By sliding the filter over the input frame, the Dot product of the filter and the input image are taken, where the resulting output is the same size asthe filter. The corners and edges of this frame are defined by the resulting output of this layer which is also called Feature Map. The output acquired is sent to the next layer.

Max-Pooling Layer: This layer is the second in the feature extraction process, and its major goal is to lower the size of the convolved feature map (the previous layer's output) in order to reduce computational costs. This is accomplished by lowering the number of connections between layers. As a result, the feature maps are controlled independently.

The largest element in this layer comes from the feature map (derived from the previous layer). The mean of the elements in an output-sized frame section is calculated using Mean Pooling. Pooling the sum computes the total sum of the components in the predefined section. The Pooling Layer

is typically used to connect the Convolutional and Fully Connected Layers.

Fully Connected Layer: It contains the biases and weights, as well as the neurons, and is used to link them between two layers. These layers are executed before the output layer and make up the final few layers of the CNN architecture. The previous layers' input image is flattened and sent to the Fully Connected layer. After that, the flattened vector proceeds through a couple additional Fully Connected layers, where the mat is added.

Activation Function: They're utilized to approximate any form of network variable-to-variable relationship that's both continuous and complex. In simple words, they decide if certain information of the model should move forward at the end of the neural network or not. Our model uses the Relu activation function as it lowers the risk of vanishing gradient.

IV. METHODOLOGY

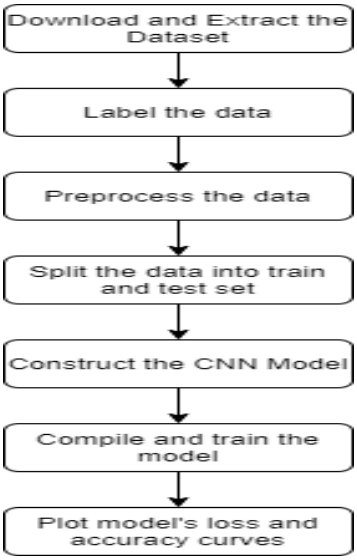

Fig.2 Implementation of Model

Extraction of the dataset to the code via python libraries and tools. Visualize the dataset with labels for a better understandingof the procedure further: Picked random classes from the dataset and labelled the respective activity to it.

Pre-processing the Data: frame extraction function: reads the video file frame by frame, resizes each frame, normalizes the resized frame, appends the normalized frame to a list, and finally returns the list. Creating a new dataset of the frames extracted. The dataset achieved is further Split the data into training and testing sets.

Construct the model: The model is classified with 2 Convolution Neural Network layers using the Relu activation function. Train andcompile the model: The model is trained with an accuracy of 90%. Plot the model's Loss and accuracy curve using python libraries and tools. Make predictions with a random video. Single framed method for Prediction: This method entails developing a function that generates a single prediction for the entire movie. This function will create predictions based on a set number of frames from the video. Finally, the mean of the forecasts of those numbers of frames will be used to determine the final activity class for that movie.

V. RESULT AND DISCUSSION

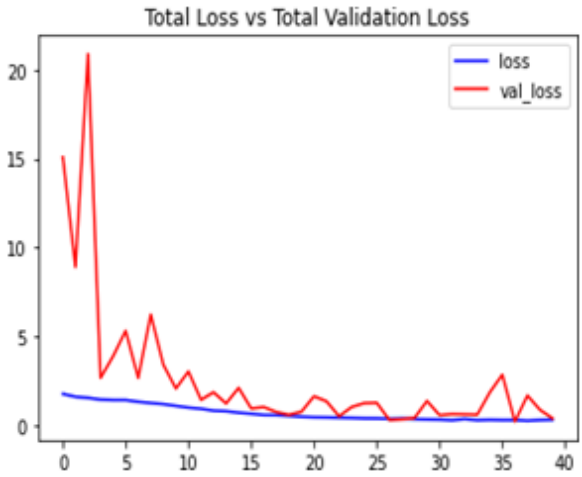

Fig.3 Loss Curves

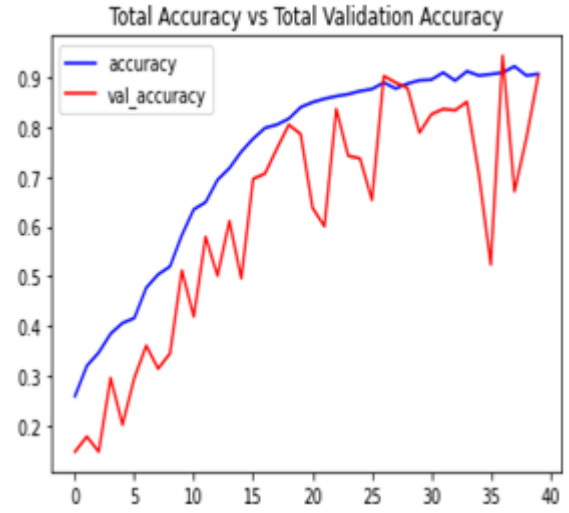

Fig.4 Accuracy Curves

```
# Constructing The Input YouTube Video Path
input_video_file_path = f'C:/Users/USER/Desktop/Capstone Project/segmented_set1/Z5_5_4.avi'

# Calling The Make Average Method To Start The Process
make_average_predictions(input_video_file_path, 50)

# Play Video File in the Notebook
VideoFileClip(input_video_file_path).ipython_display(width = 500)
```

```
CLASS NAME: Punching   AVERAGED PROBABILITY: 98.6375447511673
CLASS NAME: Hugging   AVERAGED PROBABILITY: 1.3501645739770538
CLASS NAME: Handshaking   AVERAGED PROBABILITY: 0.011555294406641714
CLASS NAME: Pushing   AVERAGED PROBABILITY: 0.0006886627678588306
CLASS NAME: Kicking   AVERAGED PROBABILITY: 5.2898117189920056-05
CLASS NAME: Pointing   AVERAGED PROBABILITY: 1.6389972226402625e-06
Moviepy - Building video __temp__.mp4.
Moviepy - Writing video __temp__.mp4
```

Fig.5 Recognition result of UT-interactiondataset

Fig.6 Recognition video playing in thenotebook

Figure 3 and 4 shows the plot of loss vs validation loss and the plot of Accuracy vs validation accuracy of the CNN model respectively. The accuracy of this model is 90.64% whereas the validation accuracy of the code is approximately 90.34%. The model was trained for 6 activities. Figure 5 successfully predicts the probability of the respective activity which was being tested. So using single frame CNN method we can predict the activities that is being performed in the video, it will average those n frame predictions and then give us the final activity class for that video in the form of likelihood. Though the probabilities are approximate, the probability of the activity tested stands out and is easily identified.

VI. CONCLUSION

In this paper, a VGG-16 and CNN-based technique for human activity recognition is proposed. To evaluate the performance of the suggested technique, extensive experiments were carried out on the UT-interaction dataset. The experimental results showed that the average accuracy of our approach was 60% and 90.64% respectively. Therefore, VGG-16 proved to be unreliable when it comes to raw video-graphic datasets. The models take more training time as observed. On the other hand CNN model proved to be very efficient for human activity recognition.

REFERENCES

[1]. Chih-Ta Yen, Jia-Xian Liao, Yi-Kai Huang, “Human Daily Activity Recognition Performed Using Wearable Inertial Sensors Combined With Deep Learning Algorithms”, Access IEEE, vol. 8, pp. 174105-174114, 2020.
[2]. Cruciani, F. Vafeiadis, A. Nugent, C.et al. Feature learning for Human Activity Recognition using Convolutional Neural Networks. CCF Trans. Pervasive Comp. Interact 2, 18-32(2020).
[3]. Semwal, V.B. (2021). Dua 2021 Article Multi input CNN GRU BasedHuman Activity Recognition.
[4]. Shugang Zhang, Zhiqiang Wei, Jei Nei, Lei Huang, Shuang Wang, Zhen Li, “A Review on Human Activity Recognition Using Vision Based Method”, Journal of Healthcare Engineering, vol. 2017, Article ID 3090343, 31 pages, 2017.
[5]. Jian Sun, Yongling Fu, Shengguang Li, JieHe, Cheng Xu, Li Tan, “Sequential Human Activity Recognition Based on Deep Convolutional Network and Extreme Learning Machine Using Wearable Sensors”, Journal of Sensors, vol. 2018, Article ID 8580959, 10 pages, 2018.
[6]. A. Deshpande and K. K. Warhade, “An Improved Model for Human Activity Recognition by Integrated feature Approach and Optimized SVM”, 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 571-576, doi: 10.1109/ESCI50559.2021.9396914.
[7]. Shreyas, D.G., Raksha, S. & Prasad, B.G. Implementation of an Anomalous Human Activity Recognition System. SN COMPUT. SCI. 1, 168 (2020).
[8]. Golestani, N., Moghaddam, M. Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks. Nat Commun 11,1551 (2020).
[9]. Vrigkas Michalis, Nikou Christophorus, Kakadiaris Ioannis A. “A Review of Human Activity Recognition Methods”, JOURNAL=Frontiers in Robotics and AI, Vol.2, Year 2015.
[10]. C. Dhiman, D.K. Vishwakarma, Engineering Applications of ArtificialIntelligence 77 (2019) 21-45.
[11]. Abdellaoui, M., Douik, A. (2020). Human action recognition in video sequences using deep belief networks. Traitement du Signal, Vol. 37, No. 1, pp. 37-44.
[12]. Liu, C., Ying, J., Yang, H. *et al.* Improved human action recognition approach based on two-stream convolutionalneural network model. *Vis Comput* **37,** 1327–1341 (2021).
[13]. “Ryoo, M. S. and Aggarwal, J. K.”, “UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities SDHA”, 2010.
[14]. J.K. Aggarwal, Lu Xia, Humanactivityrecognition from 3D data: A review,Pattern Recognition Letters, Volume 48, 2014, Pages 70-80, ISSN 0167-8655.
[15]. N. Oliver, E. Horvitz and A. Garg, "Layered representations for human activity recognition," Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, 2002, pp. 3-8, doi: 10.1109/ICMI.2002.1166960.