International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Real Time Human-Human Interaction Recognition Using Convolutional Neural Networks

Harshitha R[1], Kavya S Muttur[2], Nandini R[3], Jayanth P[4], M S Srividya[5], Anala M R[6]

[1-4]Student, Department of Computer Science & Engineering, R V College of Engineering, Karnataka, India
[5]Assistant Professor,[6]Professor Department of Computer Science & Engineering, R V College of Engineering, Karnataka, India

---

Abstract-*In this modern era, several methods have been implemented for the identification of interaction between two humans. Real time Human-Human interaction recognition using CNN is a method to identify the interaction between two humans. This method uses Mobile net SSD trained model for Human object detection, Opencv functions for Video processing, and ResNet 50 CNN model for interaction recognition and there are many applications that can be implemented by using this method. For example, we can make use of this system to identify the interacted people with covid affected person. The CNN Model is Trained with 3000+ images and tested with 800 images.*

Key Words: Human interaction, Open Source Computer Vision Library (OpenCV), Convolutional Neural Networks, Deep Learning, TensorFlow.

1.INTRODUCTION

Human interaction classification in video sequences is a fascinating and difficult subject in a variety of applications, including video surveillance. The difficulties in interaction recognition arise from differences in the basic style, speed of activity, clothes, and partial occlusions detected among video sequences featuring the same interaction.The Covid-19 epidemic has resulted in a significant loss of human life around the world and poses an unprecedented threat to public health, food systems, and, as a result, the labour market. And its very important that if one person is affected by covid then we need to find the people who was had a contact with them. Its not much easy to identify the people interacted with them if the Covid affected person was travelled when they get affected. We can expect that so many people would have interacted. Hence Real-time Human-Human interaction recognition will prove to be a handy and smarter solution to identify Such peoples. It identifies the interaction between two humans like handshake, hug, slap, kick, point and punch.

2.RELATED WORK

There are many number of approaches that are defined and implemented for human interaction recognition and we have studied some of the analysis, and these are mentioned in the following survey. In [1], Analyzing human-human interaction using two stream CNN Model was done. This survey gives a summary of the issues as well as datasets to address them. In [2], Deep learning models for the perception human social interactions using CNN model done through the two CNN architecture such as VGG 16, ResNet 50. This paper provides a summary that only deep neural networks could recognize social interactions. In [3], Human interaction recognition through deep learning network using CNN is done through using two autoencoders. The problem with this paper is model is trained with less data. In [4], 3D CNN for human action recognition will be done using 3D CNN model. The problem with this paper is it requires a large number of labeled samples. In [5], A novel approach for robust multi human action recognition and summarization based on 3D CNN is done through 3D CNN model and saving each person action at each time of the video. The problem with this paper is the accuracy of RGB videos are less. In [6], Interaction relational network for mutual action recognition will be done using relational network. The problem with this paper is its not suitable for group activity recognition. The Laplacian of Gaussian (LoG) of movies is obtained using the 3D Laplacian in[7], and trajectories are built from dense interest spots. Interest point detectors and other trajectory-based approaches are outperformed by dense trajectories [8]. Graphical Models [9] efficiently and successfully manage big amounts of movies. It captures the interdependence of several components in video sequences of the same type. In diverse interactions, the discriminative model[10] focuses on obscured body parts and visually identical individual movements.

With the help of the survey of these papers, the idea towards the concept comes into picture and this helps to choose the better model. And this survey provides overall concept and their implementations ideas and what are our challenges towards the project and what are the necessary technologies that needs to be used and in turn, provides on overall concept of our application.

3.PROPOSED SYSTEM

Here ,In this section several steps to detect and recognize the human object and interaction between those two humans in the video stream or real time video will be discussed. The complete method has been divided into different steps and the same are discussed here.

In this model, pre-process of the image is initially executed and the video input is provided for which we want the prediction to be conducted. First the video frames are

resized to the trained frame resolution and the human object detected through the mobile net model and then the final phase which would be the recognition phase where the interaction is recognized with the help of CNN will be displayed respectively.

3.1 Input the video

The initial steps that need to be taken is to load the input video using Opencv functions. For real time video we are enabling the camera interface by using Opencv functions.

3.2 Pre-Processing of the input video

This step includes the preprocessing of the input video. The video is resized into the trained resolution. The order of the color is converted. The pre-process of the video is been conducted using Opencv platform functions.

3.3 Human object detection

In this phase our aim is to display the bounding box around the interacting individuals. Such that we can easily identify the human object in the video stream. So to achieve this we will be using Mobilenet SSD Caffe model. SSD is a prominent object identification technique, and Mobilenet is a CNN that generates high-level features. Combined both the SSD and Mobilenet models to provide high accuracy. Therefore, the final output here will be a video with human identification. A basic outline of the methodology is been displayed figure-1 below.

3.4 Interaction recognition and CNN model architecture

To recognize interactions like hug, kick, slap, handshake between two humans CNN Model is used. We will be using ResNet 50 which is one of the CNN network. We trained this model to identify 6 different interactions viz hug, kick, slap, point, handshake, punch. Before sending the video to predict from the CNN model, we are resizing the frames in the video. After resizing, number of channels are chosen from the model and then sent it to prediction with Input video of 244 x 244 resolution.

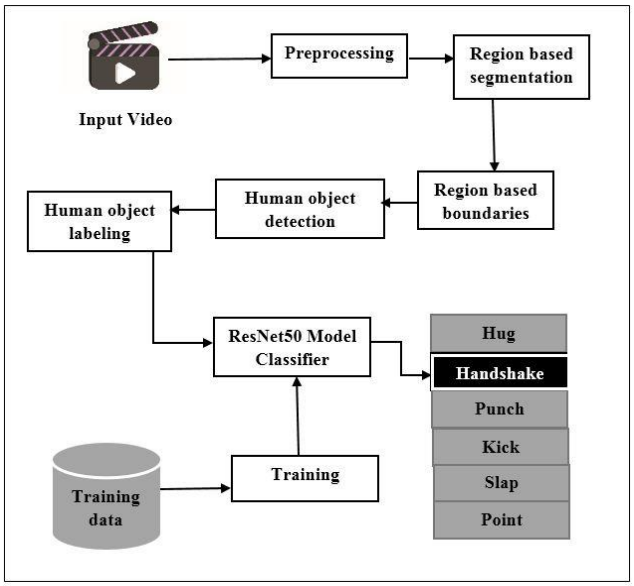

Fig-1: System architecture

3.4 Interaction recognition and CNN model architecture

To recognize interactions like hug, kick, slap, handshake between two humans CNN Model is used. We will be using ResNet 50 which is one of the CNN network. We trained this model to identify 6 different interactions viz hug, kick, slap, point, handshake, punch. Before sending the video to predict from the CNN model, we are resizing the frames in the video. After resizing, number of channels are chosen from the model and then sent it to prediction with Input video of 244 x 244 resolution.

The ResNet-50 model is divided into five stages, each with its own convolution and identity block. There are three convolution layers in each convolution block, and three convolution layers in each identity block. There are around 23 million trainable parameters in the ResNet-50.The structure of the Proposed CNN architecture is given at Table-1 below:

Table-1: CNN Architecture

Layer Type	Description
9 Convolution Layers	3 Set of filters of size [64,64,256], ReLU
12 Convolution Layers	3 Set of filters of size [128,128,512], ReLU
18 Convolution Layers	3 Set of filters of size [256,256,1024], ReLU
9 Convolution Layers	3 Set of filters of size [512,512,2048], ReLU

Average Pooling layer	Pool Size 2 X 2
Flatten Layer	Making it into a single column
First dense layer	512 X 128, ReLU
Dropout layer	Dropout:0.5
Second dense layer	128 X 6, SoftMax
Output Layer	6 Classes ranging from A-Z

The input to the network is a pre-processed 244 X 244 video. The network consists of 5 stages. In stage 1,First layer is 2D Convolution which has 64 filters of shape (7,7) with Relu activation function and max pooling layer. In stage 2,consists of 9 convolution layers uses 3 set of filters. In stage 3,consists of 9 convolution layers uses 3 set of filters. In stage 4,consists of 18 convolution layers and uses 3 set of filters. In stage 5,consists of 9 convolution layers uses 3 set of filters. The next is output layer where it consists Average pool layer, Flatten layer,2 dense layers and dropout layer. Average pooling Layer follows each of the convolution layer with the pool size of 2 X 2.Dense layer with the dimensions of 512 X 128(first dense layer) and 128 X 6(second dense layer) with the dropout having 0.5 after first dense layer were chosen with the activation functions of Relu and SoftMax for the final output.

Hyper Parameters are all the parameters which can be arbitrarily set by the user before starting the training process. The table below shows the hyper parameter set. The neural network was built using the python Keras module. Below the table-2 is the hyper Parameter table:

Table-2: Hyper Parameter Table

Hyper Parameter	**Value**
Number of Epochs	150
Loss function	Categorical Cross Entropy
Optimizer	Stochastic gradient descent
Learning rate	0.0001
Convolution Kernel Size	64 X 64
Average Pooling Kernel Size	2 X 2

4. RESULTS AND DISCUSSION

For the Experimental Results, we created our own dataset consisting of 3000+ images for training and for testing. This dataset was used in the Interaction recognition phase where the interaction between two humans would be extracted. The images is collected from various sources like From animated films, movies, internet sources, Sports videos, Events and so on.

Below at figures shows the results that we have obtained after execution.

Fig-2: Results

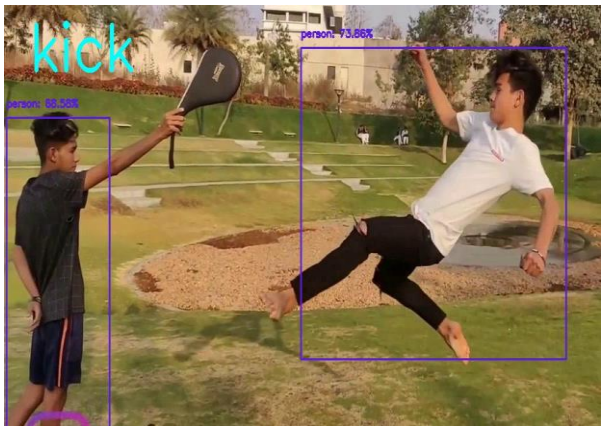

Fig-3: Output image

Fig-4:Output image

Fig-5: Results

Below chart 1 shows the increase in accuracy for each epochs. After the 150 successful epochs we got 81.42 accuracy for this model.

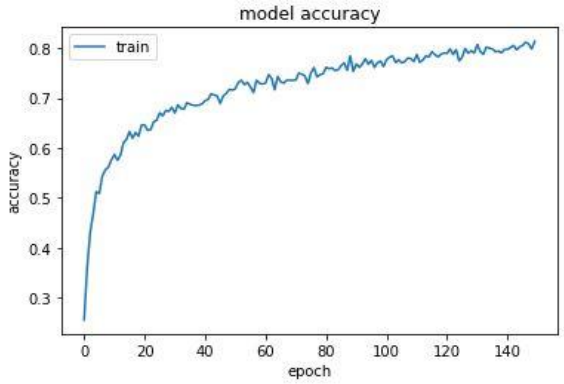

Chart-1: Model accuracy

5. CONCLUSION

In this project, we have developed a method for human detection and interaction recognition using OpenCV tools and CNN Model. The final result that we have provided is a interaction between two humans. We have made sure that the approach will give best results in Realtime and provide a research base to other researchers to carry further work in the field of image processing and deep learning. Our system could also integrate with other things like in CCTV cameras and so on. There can be further improvement that can be done in recognition phase by improving the dataset and implementing the system for more interactions.

6. REFERENCES

[1] Alexandros Stergiou, Ronald Poppe, "Analyzing human–human interactions" Computer Vision and Image Understanding 188 (2019)

[2] Elizabeth Merritt Eastman, "Deep Learning Models for the Perception of Human Social Interactions" IEEE June 2019.

[3] S. Jeba Berlin and Mala John, "Human Interaction Recognition through Deep Learning Network" IEEE October 2019

[4] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach.Intell.,vol. 35 ,pp. 221–231,January 2013.

[5] Noor Almaadeed , Omar Elharrouss, Somaya Al-Maadeed, Ahmed Bouridane,Azeddine Beghdadi, "A Novel Approach for Robust

[6] Multi Human Action Recognition and Summarization based on 3D Convolutional Neural Networks" Computer Vision and Image Understanding 188 (2019)

[7] Mauricio Perez, Member, IEEE, Jun Liu, and Alex C. Kot, Fellow, IEEE, "Interaction Relational Network for Mutual Action Recognition" 7 Jan 2021,IEEE

[8] R.D. Geest. T. Tuytelaars, "Dense interest features for video processing," in Proc. IEEE Int. Conf. Image Process.,Oct. 2014, pp.5771– 5775

[9] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, "Action recognition by dense trajectories," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., June 2011, pp. 3169–3176.

[10] H. Sahbi, "Bags-of-daglets for action recognition," in Proc.IEEE Int. Conf. Image Process., Oct.2014, pp.1150 – 1154.

[11] Y. Kong, W. Liang, Z. Dong, Y. Jia, "Recognizing human interaction from videos by a discriminative model," IET Comp. Vis.,pp. 277-286,August 2014.