International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 675-683

Interpretation of Human Actions from Live Video using Deep Learning Techniques

N. Prasath
Associate Professor
Department of Computer Science and Engineering
KPR Institute of Engineering and Technology
Coimbatore, Tamil Nadu, India
n.prasath@kpriet.ac.in

R. Abijoy
Department of Computer Science and Engineering
KPR Institute of Engineering and Technology,
Coimbatore, Tamil Nadu, India
riabijoyfs@gmail.com

Abstract

The human monitoring system is one of the greatest challenges faced by various researchers in all these years. There are various methods and techniques proposed by differerent professionals who work in this area. The activity is monitored using the camera and the detection is done using neural networks. Every method implies different results based on their characteristic features included in the model. There are different approaches including the yolov3, object detection algorithms and various neural networks and infer different successful results. To improve the additional features along with the existing model, the work is carried out with the deep learning techniques and computer vision tools and the results obtained shows a comparatively greater efficiency and can be achieved using simpler process. The execution of the process is achieved by deep learning techniques by constructing the convolutional neural networks and computer vision. This allows the model in detection and monitoring of multiple human activities from the video. The paper explains the execution of the model and the results obtained by employing this method.

Keywords: *monitoring, detection, deep learning, neural network, multiple human action*

I. INTRODUCTION

The advancements in the research have made the development process in a simple and much effective manner that is more useful in various aspects in the achievement of the greater results of all time.[1] The Deep Learning techniques are widely used in various applications such as the object detection, object recognition, motion tracking and multiple object tracking. This can be achieved by the Convolution Neural Networks built from the dataset gathered from various sources as input and functions to perform the necessary operations to produce the desired output.[2] The dataset is collected as images and videos since the model is to be trained to meet the input type, process the input and produce the output. The collection is done from various sources and various aspects since, the input cannot be the same as that of the trained data. To avoid the errors, provide optimum accuracy in the recognition and highly trained network, the duplication of the dataset is done to obtain the various forms of the input. [3]

The training model is developed using the three ways: training from scratch, training from transfer learning, training from feature extraction. The training from scratch is the method in which a new training model is developed from the scratch and the functions and the features are to be designed based on the needs of the system.[4] This is a long process but has the liberty for the developer to design the system as per the ideas of the person from the beginning. The training from the transfer learning is an effective technique in the development of a training model which can be achieved by simply modifying the existing training model by adding additional features to perform the respective calculations and operations [5]. This method is

ISSN: 2233-7857 IJFGCN

efficient and suggested in most of the cases where there is a short span of time in the development process. The training from feature extraction is an extensive method in which the features are collectively gathered from various sources of training models and are integrated together as desired by the developer to achieve the expected training model [6].

The project deals with the live video and hence the webcams are used in obtaining the input video for the process. The camera is connected through the hardware support tools and the object is detected based on the training model [7]. The object detection allows the person to classify the object and track the object throughout the scene. The live video is processed as frames and the classification of the objects is done and the action is recognized, and the output is given as the layout of the object with the classified action as the title [8]. This helps the observer to easily classify the actions of the multiple persons in a same frame. This monitoring system helps to observe the deployment area and infer the activities of the person in the scene.

The live video is processed by using various techniques by optimization of the video frames and the classification of the activity is done with the trained dataset for the activities. The training of the data is done by adding each activity as an input to the CNN model and the evaluation is done for the inputs to detect and classify the activities for the sample inputs [9]. The trained model is tested for a certain number of epochs and then the model is prepared for the interpretation process. The trained model is used to process the live video and to infer the activity performed by the person in the video. Since the video feed contains multiple persons, it is necessary to process each person as an object and the classification should be done for every person carrying out either same activity or different activities [10]. This is the major challenge faced in the process of using the live video as input and deploying in an open area. The model is trained for various activities such as walking, jogging, running, and other day-to-day activities and the detection helps to classify the actions of the people based on the time and the environment to learn their activities with time [11].

II. RELATED WORK

The Human activity monitoring can be achieved by various techniques and each method has its unique features and advantages. Convolution Neural Network is one of the widely used deep learning techniques in which various image processing and the deep learning applications are developed. Deep Learning allows the ability to train the model using large number of inputs and classify various actions of the human from the images and videos. There are various works done based on the deep learning techniques. First, object detection was done by training the model to detect, analyse the object using CNN. The neural network is designed as various layers and the model is trained to detect the objects from the video.

As a further step in the object detection from video, the motion tracking of the object is done by tracking the path of the object. The object detection and tracking are achieved by this process. The object detection is improved to detect and track human from the video and the process is done by various steps such as frame separation, optimization and the processing of the frame with the trained model to detect the human and track the movement of the human from the video. The work is improved to detect multiple humans in a single video and the tracking. The multiple human detection is done by selecting the human region with a rectangular line region in which the line is used to track the movement. The line moves along with the movement of the human throughout the video.

The detection of a single person in a video is simple but the detection of multiple human and tracking from a video is an extremely tedious process. Thus, various improvements are made through the years in the human detection and tracking from video input. This is used in the monitoring of the human activities in various environment to analyse the actions of the human with respect to the surroundings and infer the solutions to the detection of the human and the activities of the person with the various dataset such as Kth dataset, HMDB51 and UCF101 datasets and various video datasets. The Kth dataset is the beginning dataset in which only the six human actions are used to train and detect the human actions. Later, HMDB51 and UCF101 datasets are developed further with more than 100 human actions which can be used to train the

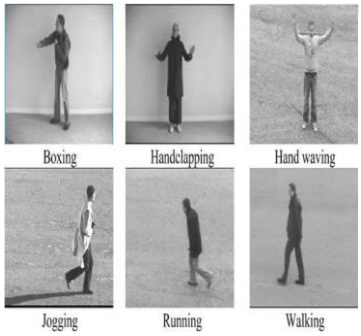
ISSN: 2233-7857 IJFGCN

model to detect the human activities. Also, various datasets are used in the training which are available online and user generated human action datasets.

III. DATASET

1. Collection of Dataset

The Dataset is collected for various human actions performed by human as a daily activity and the video of the actions is collected. There are various pre-collected datasets of human actions such as Kth dataset (Fig. 1), HMDB51 (Fig. 2) and UCF101 (Fig. 3) datasets are used for training which has over 100 human actions. The Kth Dataset consists of six major actions such as Boxing, Handclapping, Hand-waving, Jogging, Running and Walking. The HMDB51 and UCF101 consists of various video clips of human actions. The collection can be done manually by collecting the video from the human action for various new activity. It is the important process in training the model and involved in the detection of the human action.



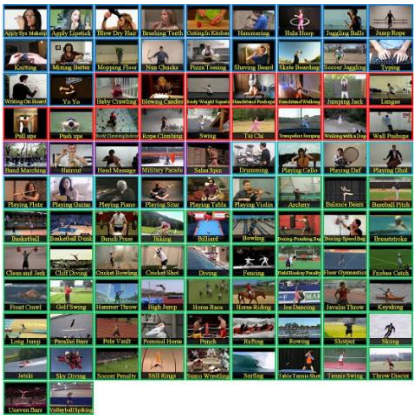


Fig. 1 Kth Dataset Fig. 2 HMDB51 Dataset Fig. 3 UCF101 Dataset

2. Analysis of Dataset

The Datasets are analysed by checking various actions in the video and their representations so that the neural network can be trained efficiently to work for different inputs. There are various customised dataset are also used along with the Human Action Datasets since the model is to be trained and to support the working of the model for different video data obtained from different environmental conditions and different person inputs. This allows the model to successfully deduce the results from the video, even though the characteristics of the video are not very clear. To ensure this the datasets are compared and the analysis is done.

IV. OVERVIEW

A. PROPOSED APPROACH

The proposed model for the monitoring process includes four phases as illustrated in Fig. 4 The architecture explaines the process involved in the construction of the model and the implementation can be achieved by this four phases. The video is captured from a camera and is processed as input for the process. The video is processed by using the human detection and the region of interest algorithms for the detection of the human from the video using the rectangular line as the region fo the detection and tracking process. The detected video is passed into the convolution neural network. The Convolution Neural Network is used in the segregation and the training of the model in detection and tracking the human from the video and the

ISSN: 2233-7857 IJFGCN

action performed is classified in the CNN. The video is used recursively to train the model and thereby improving the accuracy in the results in the process. The output generated consists of the detected human from the video and their corressponding action performed in the video. This is applied for multiple person present in the video.

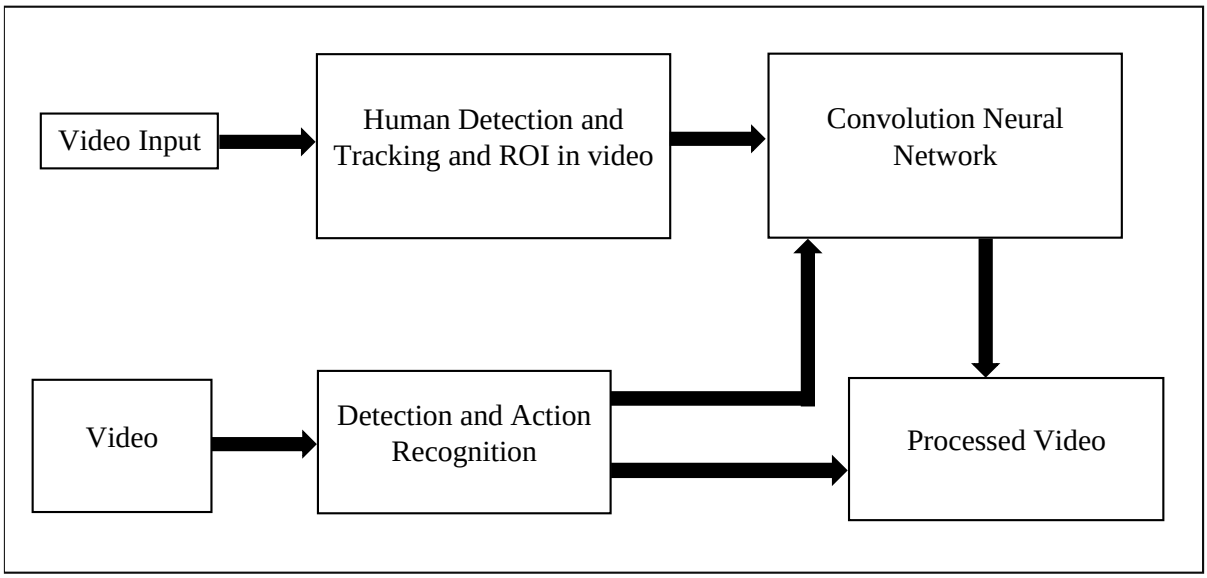


Fig. 4 Architecture of monitoring process

PRE-PROCESSING PHASE

In this process, the video is viewed to ensure it satisfies the resolution and the dimensions supported by the model. The pre-processing phase includes the human detection, tracking and the frame segmentation process. The video can be modified by processing the video input into frames by the segmentation process. The segmentation can be done by converting each frame available in the video as a sequence of images with equal time intervals. The human detection and tracking can be achieved by the people detector and the ROI (Region of Interest) algorithms which are explained in the further process.

The training and testing process are explained in the Fig. 5 as the both process are done very effectively in order to generate the optimum results. The training process can be explained as the model is trained based on the input video and the corrections in the process can be made during the training. This helps the model to support various inputs and train the model accordingly to deduce the results for different video quality and different human actions. The model is trained with different set of actions in which the model can be deployed to recognize them.

The testing process is the process to ensure that the trained model is working with the same efficiency as that of the training and in this process, the observations are made during the execution of the model. The testing is one of the important task in which the model is ensured that, it should recognize the actions for the trained videos and also some of the random videos with the trained actions. This ensures that the training is successfully carried out and the efficiency of the model is improved.

ISSN: 2233-7857 IJFGCN

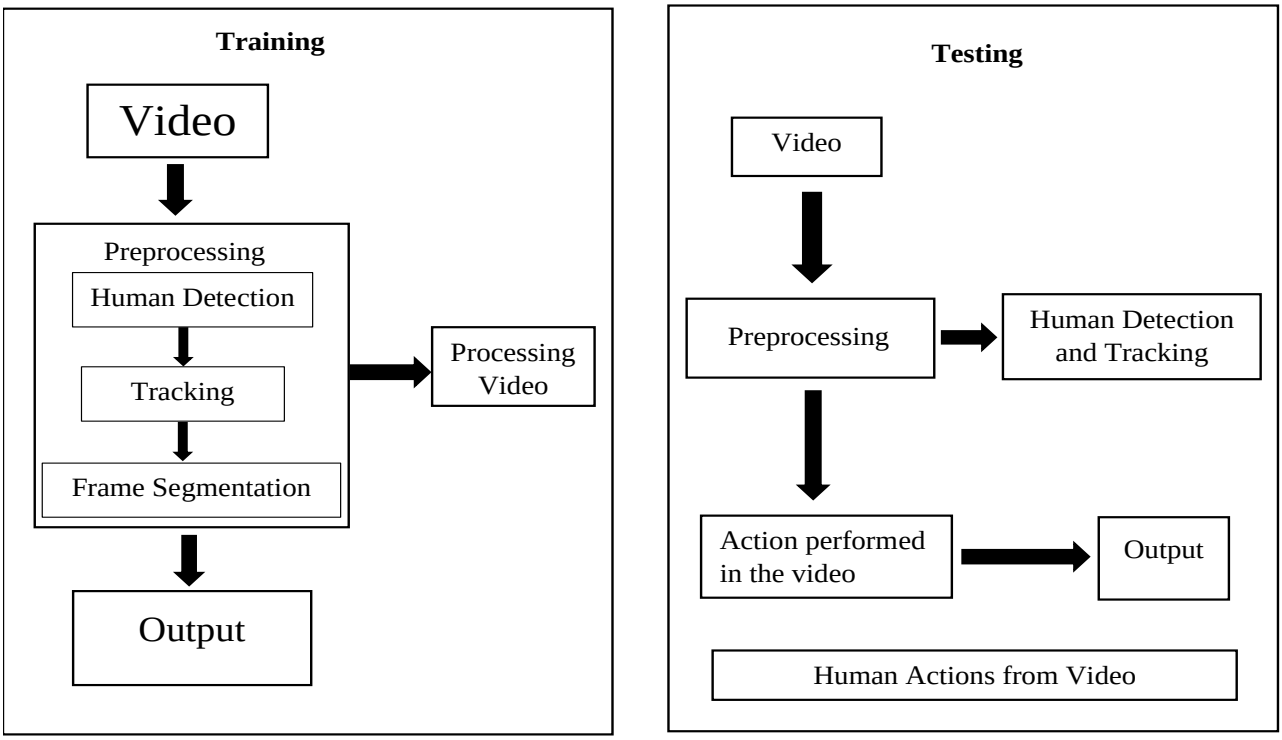


Fig. 5 Training and Testing Process

CNN MODEL

The Convolutional Neural Network is a deep learning technique in which a convolution model is generated using the neural networks. The convolution is the process in which the calculations are done in the model. The neural network is a connection of layers classified into input layer, hidden layers and output layer. This can be explained by the Fig. 6

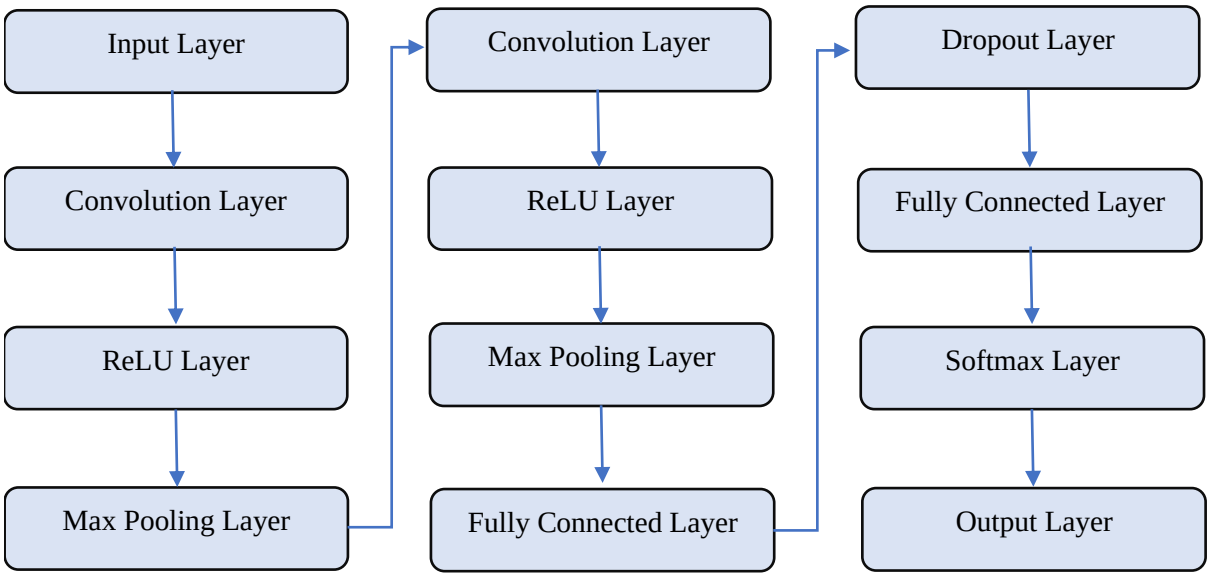


Fig. 6 Convolution Neural Network

ISSN: 2233-7857 IJFGCN

Input Layer
The input layer is the layer in which the input image sequence obtained from the video are allowed to enter. The input layer can be denoted by I(x,y) since the two dimensional image input.

Hidden Layers
The hidden layers include the Convolution layer, ReLU layer, Max Pooling layer, Dropout layer, Softmax layer and Fully Connected layer. These layers are the hidden layers in which each are used for the specified process such as the Conv(x,y,f) is the convolution layer and ReLU represents the rectified linear unit which is the activation function. The ReLU can be given by,

Softmax layer is used to identify the loss in the layers. The fully connected layer is the FC(n) with n neurons which is the classification layer. The max pooling layer is a pooling layer in which the maximum values from each pool are selected and given as Mpool(x,y,k). It can be calculated by the formula,

Output Layer
The output layer is the layer in which the action performed is generated as output and it is given along with the video as the labels with the person and the corresponding action performed.

B. TRAINING THE MODEL
The model is designed using the CNN in order to detect the human actions from the live video. The Computer Vision, Image Processing Tools and the Groung Truth Labeler are used to support the process. The Convolution Neural Network is designed as three layers such as input layer, hidden layer, and output layer. The Input layer is the layer in which the inputs are pre-processed for the subsequent layers, the hidden layer is the training layer of the CNN and the output layer is the classification layer.

1. Labeling the human using video labeler
The ground truth labeler is tool is used to label the human from the video to train the model to detect the human. The person in the video is selected using the ROI (Region of Interest) in rectangular shape to set the boundary for the person in the video. The region of interest of person is selected in each frame from the video. This helps the labeler to help the process. The ACF People Detector algorithm is selected for automation of the model to detect human. This can be done by running the automated model. The detection results can be modified for corrections and the data is generated.

2. Exporting the gTruth data to workspace
The labeling is done and the values are generated from the video labeler. The generated data values are exported to the file or directly to the workspace as gTruth. The gTruth is a variable which consists the numerical values of the people labeled in the labeler. It is used to train the detector to detect the human from the input video. This numerical values are categorized as the people position in the frame based on the corners of the rectangular region of interest.

3. Training the human detector
The detector is trained using the simple lines of code in the mat file and the code is executed to train the detector. The programming include the layers of the model to process the exported gTruth data to the workspace and the numerical values along with the input video is used to train the model and the trained data is exported into a variable. The exported variable data can be loaded again in the workspace to detect the human in the video. The model can be tested with various sample videos in which the people are detected.

ISSN: 2233-7857 IJFGCN

4. Training the action detector

The action detector is used in the detection of the human action and it can be achieved by the trained convolutional neural network. The convolutional neural network can be designed with the various layers such as the input layer, relu layer, softmax layer, fully-connected layer and output layer. These layers are the part of the nueral network used in training the network. The input layer is used to provide the input to the network. The relu layer, softmax layer aand the fully-connected layers are the intermediate layers in which the training is done. The training process is done by feeding the data in a recursive methos and the model is trained to detect the actions from the video. The actions can be detected by comparing the frames in the video with the dataset with action videos and the output is given in the output layer. The frame segmentation process is explained in the working process in the further section.

C. WORKING

1. Detection of human in video

The frames obtained from the video are used to detect the human in the frame. The region of the human is marked by a rectangular border to define the human detection. The rectangular line moved along with the movement of the human throughout the frame area. The time of validation of the region is till the presence of the person in the video frame. Since the model should detect the multiple humans in the frames, there are individual rectangular region is marked for each human in the frame. This helps the viewer to identify the human and track the movement if any within the frame.

The detection of the human from the video obtained by using ACF People Detector algorithm. ACF classification model, specified as the comma-separated pair consisting of 'Model' and either 'inria-100x41' or 'caltech-50x21'. The 'inria-100x41' model was trained using the INRIA Person dataset. The 'caltech-50x21' model was trained using the Caltech Pedestrian dataset. The ACF (Aggregate Channel Feature) People Detector is an algorithm is used to detect the people in the video. This algorithm is used also by making some user defined changes to support and enhance the process of detection. This reduces the complex steps in the process of the human detection and completes the task with a short time by using the video labeller tool from MATLAB by defining the Region of Interest (ROI) Label Definition. The automation of the algorithm is used to train the model to detect human.

2. Processing the video frames

The video is used as an input in the frame segmentation and the process is carried out by the image processing mechanisms by separating the frames from the video and the frames are processed as images. Each frame is processed so that the quality of the frame is maintained throughout the video. The frames are then proceeded further to the process of selection of region.

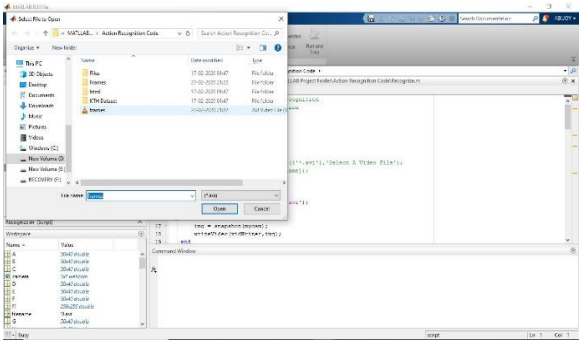


Fig. 7 Selecting the video

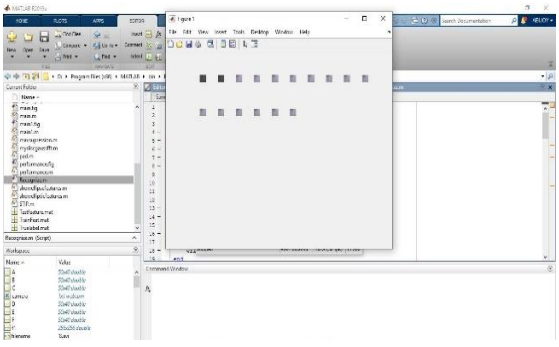


Fig. 8 Segmentation of video frames

ISSN: 2233-7857 IJFGCN

The video is selected from the saved location of the captured live video and used as shown in the Fig. 7. The video is selected and is proceeded to the segmentation of the frames process. The segmentation process is carried out as shown in the Fig. 8.

The segmentation process segments the video into 50 single frames. The process is done by generating the snapshot of the images from the video as the series of frames. The snapshot is collected by using a loop in the line of code to capture images to the size of 50. The capturing includes a sequential image without break in the time in-between each subsequent frame. The frames are named sequentially from 1 to 50 and stored in the sub-directory of the project folder named Frames. The frames are used as an input in the model to detect the action performed by the person in the video. The process is explained in the next step.

3. Inference of human action

The detection of the human action is based on the results received from the previous labelling of the human from the video labeller. This detects the human in the video. The segmentation process provides the necessary frames to detect the human action. The trained model is used with the images as input and the inference provides the result as the action performed by the person the video as shown in the Fig. 9.

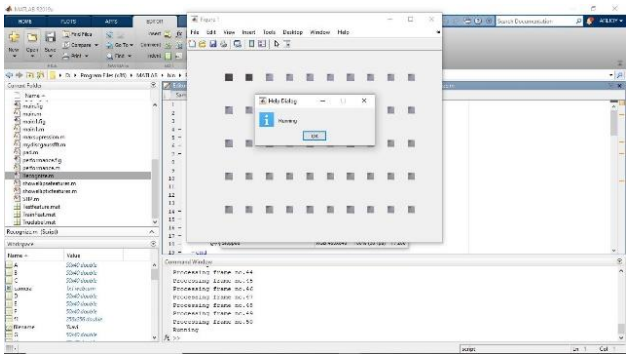


Fig. 9 Human Action Inference

The action inferred is obtained from the trained set of action which matches the action of the person from the video. The training of the action is done by using the numerical values in the files to train the model. This increases the efficiency in the results in reduced time. The inference is obtained as the action of the human.

V. CONCLUSION

The project can be further improved with various actions by training for more actions based on the deployment environment. This helps to monitor human activities in regular basis and can be used to understand the behaviour of a person from time to time to improve the efficiency in the work and reduce the stress issues faced by the major software professionals and ensure a sound mind and health of the people to maintain a healthy work environment.

References

[1] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE, 2005..
[2] Anitha, A., J. Gayatri, and K. Ashwini. "Automatic recognition of object detection using Matlab." *International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)* 2 (2013): 749-756.
[3] Malavika, T., and M. Poornima. "Moving object detection and velocity estimation using MATLAB." *International Journal of Engineering Research & Technology (IJERT) Vol* 2 (2013): 2278-0181.

ISSN: 2233-7857 IJFGCN

[4] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[5] Kumaran, N., U. Srinivasulu Reddy, and S. Saravana Kumar. "Multiple Action Recognition for Human Object with Motion Video Sequence using the Properties of HSV Color Space Applying with Region of Interest." International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-6 (2019).
[6] Almaadeed, Noor, et al. "A novel approach for robust multi human action detection and recognition based on 3-dimentional convolutional neural networks." *arXiv preprint arXiv:1907.11272* (2019).
[7] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767* (2018).
[8] Fu, Jun, et al. "Dual attention network for scene segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
[9] Husain, Farzad, Babette Dellen, and Carme Torras. "Scene understanding using deep learning." *Handbook of Neural Computation*. Academic Press, 2017. 373-382.
[10] Mehta, Sachin, et al. "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
[11] Dai, Angela, and Matthias Nießner. "3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
[12] Chen, Liang-Chieh, et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs." *arXiv preprint arXiv:1412.7062* (2014).
[13] Jin, Cheng-Bin, et al. "Real-time human action recognition using CNN over temporal images for static video surveillance cameras." *Pacific Rim Conference on Multimedia*. Springer, Cham, 2015.
[14] Zhang, Shugang, et al. "A review on human activity recognition using vision-based method." *Journal of healthcare engineering* 2017 (2017).
[15] Zhao, Hang, et al. "Hacs: Human action clips and segments dataset for recognition and temporal localization." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.
[16] Carreira, Joao, et al. "A short note on the kinetics-700 human action dataset." *arXiv preprint arXiv:1907.06987* (2019).
[17] Beale, Mark Hudson, Martin T. Hagan, and Howard B. Demuth. "Neural network toolbox™ user's guide." *The MathWorks* (2010).
[18] Vrigkas, Michalis, Christophoros Nikou, and Ioannis A. Kakadiaris. "A review of human activity recognition methods." *Frontiers in Robotics and AI* 2 (2015): 28.
[19] Zhang, Yifei, Wen Qu, and Daling Wang. "Action-scene model for human action recognition from videos." *AASRI conference on Computational Intelligence and Bioinformatics*. Vol. 2. 2014.
[20] Jin, Cheng-Bin, et al. "Real-Time Action Recognition Using Multi-level Action Descriptor and DNN." *Intelligent Video Surveillance*. IntechOpen, 2018

ISSN: 2233-7857 IJFGCN