Action Recognition based on Pose Estimation

Bachelor Thesis

By

Anshu Kumari
(39/CSE/17019/217)

A thesis submitted to

Indian Institute of Information Technology Kalyani

for the partial fulfillment of the 4th Year 8th semester Project on

Suspicious Human Behaviour Detection
in
Computer Science and Engineering

May, 2021

Certificate

This is to certify that the thesis entitled “**Multi-Person Action Recognition based on OpenPose**” being submitted by **Anshu kumari**, an undergraduate student (ID: 39/CSE/17019/217) in the Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, West Bengal 741235, India, for the award of **Bachelors of Technology** in **Computer Science and Engineering**, is an original research work carried by her under my supervision and guidance. The thesis has partially fulfilled the requirements as par the regulation of **Indian Institute of Information 0Technology, Kalyani**. The work, techniques and the results presented have not been submitted to any other university or Institute for the award of any other degree or diploma.

(**Dr. Oishila Bandyopadhyay, Ph.D**)
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Information Technology Kalyani
Kalyani, W.B.-741235, India.

Acknowledgments

First of all, I would like to take this opportunity to thank my supervisor Dr. Oishila Bandyopadhyay. Without whose effort this thesis would not have been possible. I am so grateful to her for working tirelessly after me, answering my doubts whenever and wherever possible. I am most grateful to Department of Computer Science and Engineering, IIIT Kalyani, India, for providing me this wonderful opportunity to complete my bachelor thesis.

And last but the biggest of all, I want to thank my parents, for always believing in me and letting me do what I wanted, but keeping a continuous check that I never wandered off the track from my goal.

Anshu Kumari
ID No.: 39/CSE/17019/217
Department of Computer Science and Engineering
Indian Institute of Information Technology Kalyani
Kalyani, W.B.-741235, India.

Abstract

Suspicious human activity recognition from surveillance video is an active research area of image processing and computer vision. Moreover, recognizing human activities from video sequences or still images is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and appearance. Many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. In this paper, I proposed a deep learning approach to classify human activities by tracking multiple humans in real-time scenarios or through videos. This paper consists of mainly 4 different types of human activities such as sitting, standing, running, falling-down. In general, I have discussed all the steps those have been followed to recognize the human activity from the surveillance videos in the literature; such as pose estimation using OpenPose algorithm, individuals human tracking using DeepSort framework, feature extraction, classification; activity analysis and recognition. The objective of this paper is to recognize the suspicious human activities such as running and falling-down in surveillance videos, mostly applicable on shopping mall.

Keywords: CNN, OpenPose, Sort, DeepSort, MOT, COCO

Contents

List of Acronyms

SORT	Simple Online and Real-time Tracking
DeepSORT	Simple Online and Real-time Tracking with a Deep Association Metric
MOT	Multiple Object Tracking
CNN	Convolutional Neural Network
COCO	Common Objects in Context

Chapter 1

Introduction

Human activity recognition is a broad field of study concerned with identifying the specific movement or action of a person based on sensor data. Movements are often typical activities performed indoors, such as walking, talking, standing, and sitting. They may also be more focused activities such as those types of activities performed in a kitchen or on a factory floor. The sensor data may be remotely recorded, such as video, radar, or other wireless methods. Alternately, data may be recorded directly on the subject such as by carrying custom hardware or smart phones that have accelerometers and gyroscopes.

In our daily life, we see the thousands of humans walk on roads, shopping malls, and other public areas. They intentionally or unintentionally keep interacting with each other. They also make decision on where to go and how to reach their destinations. So their movement is not always straight away. It changes based on external environmental factors. Study and analysis of human dynamics play an important role in public security, public space management, architecture, and design. These tasks are highly dependent upon proper multi-person tracking and trajectory extraction procedure. So this thing motivated us to contribute in the development of such system which performs these tasks with real-time speed and high accuracy.

1.1 Scopes of Human Activity Recognition

The main goals of Human Activity Recognition systems are to observe and analyze human activities and to interpret ongoing events successfully. Using visual and non-visual sensory data, the systems retrieve and process contextual (environmental, spatial, temporal, etc.) data to understand the human behavior. There are several application domains where these concepts are investigated and the systems are developed. Human Activity Recognition systems have some pretty cool applications and is heavily used in different fields:

- **Healthcare monitoring applications:** Healthcare monitoring systems are designed based on the combination of one or more AR components such as fall detection, human tracking, security alarm and cognitive assistance components. Once help is needed, the system notifies the relevant parties (i.e. medical personnel) about the situation to assist the patient quickly.

- **Security and surveillance applications:** The surveillance system is able to recognize human behavior such as fighting and vandalism events occurring in a metro system using one or several camera views. Additionally, this system was able to detect and predict the suspicious and aggressive behaviors of a group of individuals in a prison. The researchers used multiple camera views to detect situations such as loitering, distinct groups, and aggression scenarios in real time and in a crowded environment. The airport surveillance system proposed by is able to recognize 50 types of events including complex activities such as baggage unloading, aircraft arrival preparation, and refueling operation.

- **Visual systems in sports and outdoors:** Computer vision techniques can also be used to recognize sport activities to enhance the performance of players and analyze the game plan. It is also used to analyze an entire playground of team activities of a handball game.

1.2 Roadmap of the Thesis

The structure of the thesis is as follows:

1. The Chapter 1 is an introductory part which discusses about the Human Activity Recognition, their scopes and the roadmap of the thesis.

2. The Chapter 2 discusses the various existing methods for recognising human activities. This chapter also discusses the various deep learning approaches which are used to detecting and tracking humans and thus recognising suspicious human behaviour.

3. The Chapter 3 discusses the different types of challenges are generally faced in recognising human activities through artificial intelligence in different types of situations.

4. The Chapter 4 discusses the proposed method which I have implemented for the human activity recognition system.

5. The Chapter 5 discusses the datasets used and experimental results of the human activity recognition system.

6. The Chapter 6 comprises of the conclusion and future work of the thesis.

Chapter 2

Related Work

Human action recognition is a standard Computer Vision problem and has been well studied. The fundamental goal is to analyze a video to identify the actions taking place in the video. Essentially a video has a spatial aspect to it, the individual frames and a temporal aspect, the ordering of the frames. Some actions such as standing, running, etc can probably be identified by using just a single frame but for more complex actions such as walking vs running, bending vs falling might require more than 1 frame's information to identify it correctly. Local temporal information plays an important role in differentiating between such actions. Moreover, for some use cases, local temporal information isn't sufficient and you might need long duration temporal information to correctly identify the action or classify the video.
The problem of action recognition in videos can vary widely and there is no single approach that suits all the problem statements. In this chapter, I will briefly discuss upon a few approaches to get a sense of the existing research in this field. Traditional approaches to action recognition rely on object detection, pose detection, dense trajectories, or structural information.

Existing activity detection methods based on computer vision can be divided into two categories:

- Traditional features based
- Depth features based

2.1 Traditional Features Based

In this approach, the method [**?**] extracts four features from the bounding box around the human silhouette to describe a fall. These features were aspect ratio, fall angle, centre speed, and head speed. Then, SVM (Support Vector Machine) was used as the classifier. However, the major drawback is the inadequate description of human motion by merely using a bounding box. To overcome this problem, Gunale and Mukherji [3] utilized an ellipse to fit the physical characteristics of a person and used KNN (K-NearestNeighbor) for classification. The accuracy of this approach was 95%. The main problem of this approach is that the ellipse can only describe the external features of the human body while some local features are neglected.This approach [8] combined the elliptical model with the projection histogram along the elliptical axis to complete the extraction of the target specific pose. They achieved a high fall detection rate of 97.08%. Similarly, an other researcher [9] detected fall events based on measuring a temporal variation of pose change and body motion. Features of centroid velocity, head-to-centroid distance, the histogram of oriented gradients, and optical flow were computed. It got an accuracy of 90.6%. Lotfi used the background subtraction to extract the moving human target and then extracted the external contour, ellipse, centroid, and other characteristics of the human body. Finally, these features were fed into a MLP for fall detection. The test results show a high sensitivity of 99.52% and a high specificity of 97.38% on the UR Fall Detection Dataset [5].

2.2 Depth Features Based

With the development of deep learning in recent years, many human posture detection models based on Convolutional neural networks (CNN) have been proposed. These models can extract the key part positions of the human body, and it provides a new approach for fall detection. A scientist [6] adopted DeeperCut (a 152-layer Residual neural network) to extract the skeleton coordinate from the 2D image. The skeleton information composed of 14 key points, which can depict the outline of the central part of the human body. Then a recurrent neural network (RNN) with long and short term memory (LSTM) state units were used to process the key-point sequence. This method was suitable for the classification of video sequences and finally achieved an accuracy of 88.9%. Huang, an another scientist [4] used the OpenPose model to get 18 key points of the human body skeleton. Then a VGG-16 network was used for feature extraction and representation. There was no specific accuracy of this method on public datasets.

A brief literature survey shows that there has been a plenty of research in the area of video analysis and human action recognition. We have come a long way in the part 5-6 years after the advent of neural networks. Initially CNNs applied frame by frame helped in improving the accuracies as compared to the manual feature extraction techniques. Later 3D-CNNs further improved the accuracies of CNNs by processing multiple frames at a time. More recent architectures started focussing on RNNs and LSTMs to factor in the temporal component of the videos. Most recent architectures started incorporating attention mechanism to focus on the salient parts of the videos.

Human action recognition is still a very active research area and new approaches are still trying to solve the issues with the current approaches. Some of the existing issues are background clutter or fast irregular motion in videos, occlusion, view point changes, high computational complexity and responsiveness to illumination changes. We discuss the issues in detail in the next chapter.

The proposed approach in this paper combines the feature extraction schemes of the above two methods, which attempts to overcome the existing shortcomings.

Chapter 3

Challenges in Action Recognition System

With recognition of human behavior several forms of computer learning have been used. However, many technological challenges still face this area. Many of the problems are associated with other areas of pattern recognition, including machine vision and analysis of natural languages, while others are specific to sensor-based behavior recognition. Below are few examples of issues that the recognition group will tackle.

- **Accuracy:** Dependency of tracking accuracy on detection accuracy.
- **Occlusion:** Loses track of humans in case of Occlusion. In object tracking, occlusions significantly undermine the performance of tracking algorithms. Unlike the existing methods that solely depend on the observed target appearance to detect occluders. As we can see, in the above figure (left), the man in the background is detected, while the same guy goes undetected in the next frame (right). Now, the challenge for the tracker lies in identifying the same guy when he is detected in a much later frame and associating his older track and features with his trajectory.
- **Visual Similarity:** Overlapping bounding boxes with low Euclidean distances are associated without taking into consideration the visual similarity between them.
- **Annotating Training Dataset:** Difficult to collect data about any emergent and unpredictable events. Wide annotated data samples are needed for training and assessment of learning techniques. However, gathering and annotating sensory experience data is costly and time intensive. Annotation scarcity thus poses a major obstacle to understand sensor behavior. Furthermore, it is especially difficult to collect data about any emergent or unpredictable events (for example, accidental falling).
- **High-end system configuration:** GPU is required for better results in real-world applications.

Chapter 4

Proposed Method and its Architecture

In this chapter, we are going to discuss the method's and its architecture in detail and why it is so amazing compared to other methods. First, all persons in the observed frames are detected and tracked with the coordinates of their body keypoints being extracted meanwhile. A keypoints vectorization method is exploited to eliminate irrelevant information in the initial coordinate representation. Then, the observed keypoint sequence of each person is input to the pose prediction module adapted from sequence-to-sequence(seq2seq) architecture to predict the future keypoint sequence. Finally, the predicted pose is analyzed by the falls classifier to judge whether the person will fall down in the future. The pose prediction module and falls classifier are trained separately and tuned jointly using Le2i dataset, which contains 191 videos of various normal daily activities as well as falls performed by several actors. The contrast experiments with mainstream raw RGB-based models show the accuracy improvement of utilizing body keypoints in falls classification. Moreover, the precognition of falls is proved effective by comparisons between models that with and without the pose prediction module.

In this paper, I have described a systematic method to recognize human activities in real time using Openpose and DeepSort framework. This approach is based on the images that are captured in real time by connecting the camera and fetching the timed screenshots.

4.1 Pose estimation - OpenPose

OpenPose [1] is an open source human pose estimation library. It detects the human body key points, facial expression and positions, hand and foot key point extraction. The pretrained OpenPose model can give 15, 18, and 25 key descriptors for a human body. OpenPose model is trained with COCO dataset to extract 18 body key point coordinates.
Input images are read from pre recorded videos or cameras. Openpose uses a neural network which returns a tensor containing 57 matrices. It outputs heatmaps and Part Affinity Fields. The output tensor is a concatenation of these 2 fields.

- **Heatmaps:** Each heatmap stores a matrix which contains the confidence that a pixel contains body part. The skeleton shows 18 heatmaps are associated with each one of the body parts. The location of each body part is extracted with this 18 matrices.

- **Part Affinity Fields:** Part Affinity Fields are matrices which contains position and orientation of pairs. For every keypoint there is a PAF in x direction and one in the y direction. There are 38 such matrices which forms the skeleton of a person.

Next layer is Non Max Suppression layer(NMS). It gives the certainty for the heatmap confidence obtained for each body part. In other words, we need to extract local maximums out of a function.
As we have found out the coordinates of the body parts, we need to join them to form skeletons. Bipartite graph connects the neck and body candidates. The vertices are body parts and association between them is represented by connection candidates.

4.2 Human Tracking - DeepSort

It [7] is an extension to SORT (Simple Online and Real-time Tracker) Framework [**?**].

SORT is a simple framework to track persons in real time. It utilizes the Kalman filter features on input frames. Hungarian algorithm is employed to find the association in visual tracks. We will see the Kalman filter and Hungarian algorithm later in details. Their proposed system is only applicable for human tracking in different appearance scenes.

But SORT has some following limitations:

- It cannot track occlusion of humans for a given sequence of frames.
- It has redundancy in the number of identity switches.
- It is not as much as efficient in terms of tracking the humans.

To overcome the above drawback, the another tracking system, DeepSORT comes into the picture. It stands for Simple Online and Real-time Tracking with a deep association metric. It utilizes apparent features extracted from deep convolution neural network (CNN) for tracking the individuals. Deep SORT generates a cost matrix based on motion information and appearance features to avoid missing tracking because of occlusion or missed detection of persons. Their system includes convolution neural network for a person's apparent features trained on person re-identification dataset. DeepSORT has a high rate of missed detection for elevated view, crowded view, and distant view because detected humans are obtained from pre-trained models of object detection.

Steps involved in DeepSORT Algorithm:

- **Kalman Filter:** It is a crucial componennt in DeepSORT. It contains mainly 8 variables; (u,v,a,h,u',v',a',h') where (u,v) are centres of the bounding boxes, a is the aspect ratio and his the height of the image. The other variables are the respective velocities of the variables.These variables have only absolute position and velocity factors, since we are assuming a simple linear velocity model. The Kalman filter helps us factor in the noise in detection and uses prior state in predicting a good fit for bounding boxes.

  For each detection, we create a collection of IDs for each bounding box, that has all the necessary state information. It also has a parameter to track and delete IDs that had their last successful detection long back, as those humans would have left the scene. Also, to eliminate duplicate tracks, there is a minimum number of detection threshold for the first few frames.

- **The assignment problem:** Now that we have the new bounding boxes tracked from the Kalman filter, the next problem lies in associating new detections with the new predictions. Since they are processed independently, we have no idea on how to associate current tracking frame with incoming detection frames.

  To solve this, we need two things: A distance metric to quantify the association and an efficient algorithm to associate the data.

- **The distance metric:** Now, We evaluate the squared Mahalanobis distance (effective metric when dealing with distributions) to incorporate the uncertainties from the Kalman filter. Thresholding this distance can give us a very good idea on the actual associations. This metric is more accurate than say, euclidean distance as we are effectively measuring distance between any two distributions.

- **The efficient algorithm:** In this case, we use the standard Hungarian algorithm, which is very effective and a simple data association problem. Well, we have an human detector giving us bounding boxes, Kalman filter tracking it and giving us missing tracks, the Hungarian algorithm solving the association problem.
  Now, we use the deep learning model. Despite the effectiveness of Kalman filter, it fails

in many of the real world scenarios we mentioned above, like occlusions, different view points etc. So, to improve this, Deep sort use another distance metric based on the appearance of the object.

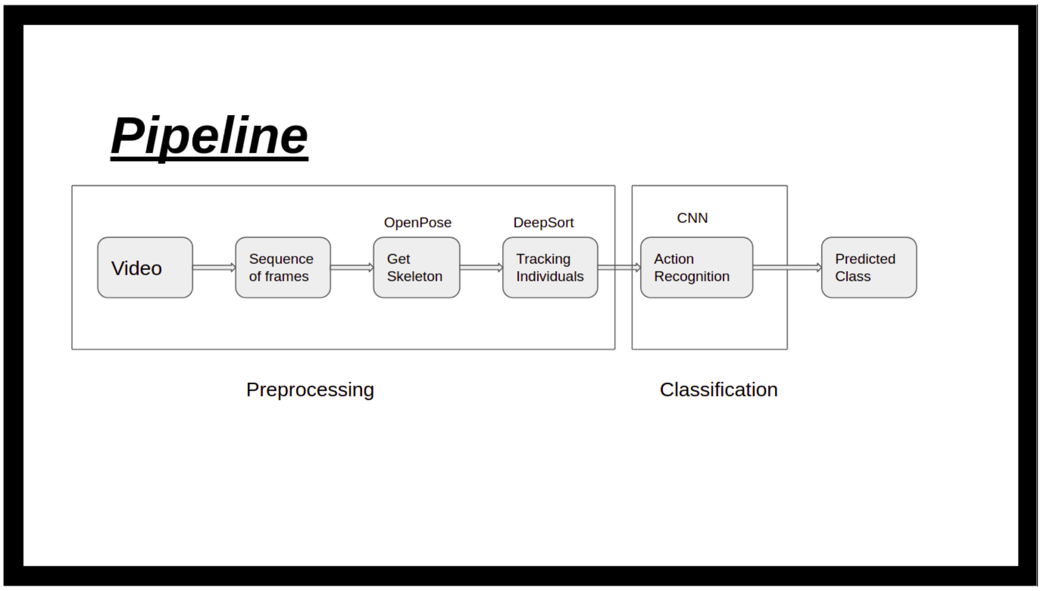
- **The appearance feature vector:** The idea to obtain a vector that can describe all the features of a given image is quite simple. We first build a classifier over our dataset, train it till it achieves a reasonably good accuracy, and then strip the final classification layer. Assuming a classical architecture, we will be left with a dense layer producing a single feature vector, waiting to be classified. That feature vector becomes our **appearance descriptor** of the object.

A simple distance metric, combined with a powerful deep learning technique is all it took for deepSORT to be an elegant and one of the most widespread human trackers.

4.3 Model Architecture

This paper proposes a method for the four different types of pose detection in video. The process of the framework is illustrated in Figure 1. This method includes three main steps: Preprocessing, Training, Classification. First, OpenPose DeepSort are used for preprocessing. The human body is subtracted from the background and then tracked in successive frames. Meanwhile, the position of the human body is also extracted. Second, the rate of centroid drop, the speed of the upper limbs, the location of key points, and the ellipse parameters of the human body are computed. These features are further divided into types of dynamic and static, which are used for describing the human body during the activity detection. Two types of features are recognized in a dual-channel sliding window of successive frames of video. Third, four classifiers are used to classify the four types of features independently. At last, the final activity detection result is judged by merging the above four classifiers.

The Pipeline the CNN model architecture of the Action Classifier of the proposed method is shown in the given figure:



Pipeline

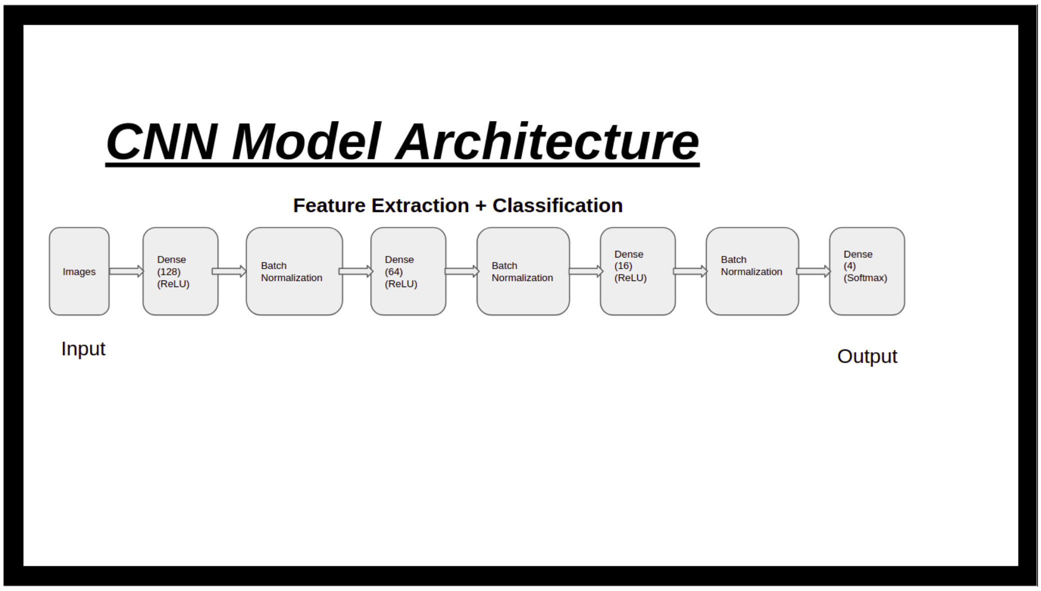
OpenPose DeepSort CNN

Video → Sequence of frames → Get Skeleton → Tracking Individuals → Action Recognition → Predicted Class

Preprocessing

Classification

Figure 4.1: The Pipeline of the proposed method.



CNN Model Architecture

Feature Extraction + Classification

Images → Dense (128) (ReLU) → Batch Normalization → Dense (64) (ReLU) → Batch Normalization → Dense (16) (ReLU) → Batch Normalization → Dense (4) (Softmax)

Input

Output

Figure 4.2: The CNN model architecture of the Action Classifier.

Chapter 5

Experiments and Results

5.1 Data Description

In order to prove the validity of our method, experiments were conducted on two public datasets: Le2i Fall Detection Dataset, and UR Fall Detection Dataset [5].

- The **Le2i Fall Detection Dataset** [2] consists of 191 human activity videos in four scenes: Office, Home, Coffee Room, and Lecture Room. The format of the video is 320*240, 25frame/s.
- The **UR Fall Detection Dataset** [5] consists of 70 depth video and corresponding RGB video (the image format is 1920*1080), which contains 30 fall videos and 40 daily activity videos from different angles and different lighting conditions.

During the training, we process the original dataset to fit our model. 23160 frames are selected from the dataset and preprocessed. According to the length of the window, the generated data vector are grouped. The dataset was relabeled, with the group containing the four events, sitting, standing, running, falling-down. The generated dataset is taken as our training dataset.

5.2 Performance Metrics

When detecting a video sequence, four possible cases are corresponding to four valid parameters (if consider for falling-down category):

- True positive (TP): a video segment contains a fall, and is correctly detected as a fall.
- False positive (FP): a video segment does not contain falls, but is incorrectly detects as a fall.

- True negative(TN): a video segment does not contain falls, and is correctly detects as non-fall.
- False negative(FN): a video segment contains a fall, but is incorrectly detected as not a fall.

Based on these four parameters, five indicators, including sensitivity, specificity, accuracy, precision and F-score, are defined to measure the reliability of fall detection method, which are calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{17}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$Precision = \frac{TP}{TP + \text{FP}} \tag{19}$$

$$F-Score = \frac{2TP}{2TP + \text{FP} + \text{FN}} \tag{20}$$

In the following classifier selection and model evaluation, these five indicators will be used as performance indicators. Among them, Sensitivity describes the sensitivity of the model to detect falls. Specificity describes the ability of the model to prevent misjudgment. The higher this index is, the lower the probability that the model misjudges other behaviors as falls. These two parameters can show the characteristics of the classification model more intuitively, so they are used as a relatively more important reference standard in the selection process of the classifier.

5.3 Testing Environment

In terms of hardware, the experiment runs on a server with Intel Core i5 CPU, and 8 Giga Bytes RAM. In terms of software, the operating system is Ubuntu 20.04, and the development language is Python 3.7.

5.4 Results

Some of the obtained outputs from the proposed methods are as shown below:

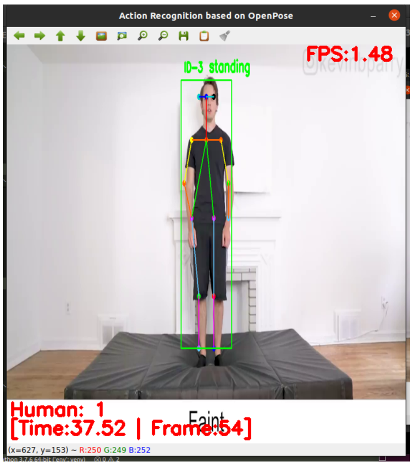


Figure 5.1: Standing Pose

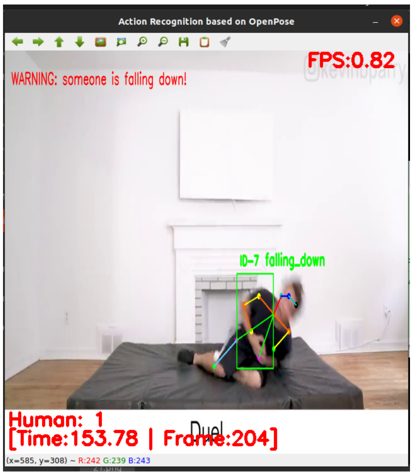

Action Recognition based on OpenPose
FPS:0.82
WARNING: someone is falling down!
ID-7 falling_down
Human: 1
[Time:153.78 | Frame:204]
Duel

Figure 5.2: Falling-down Pose

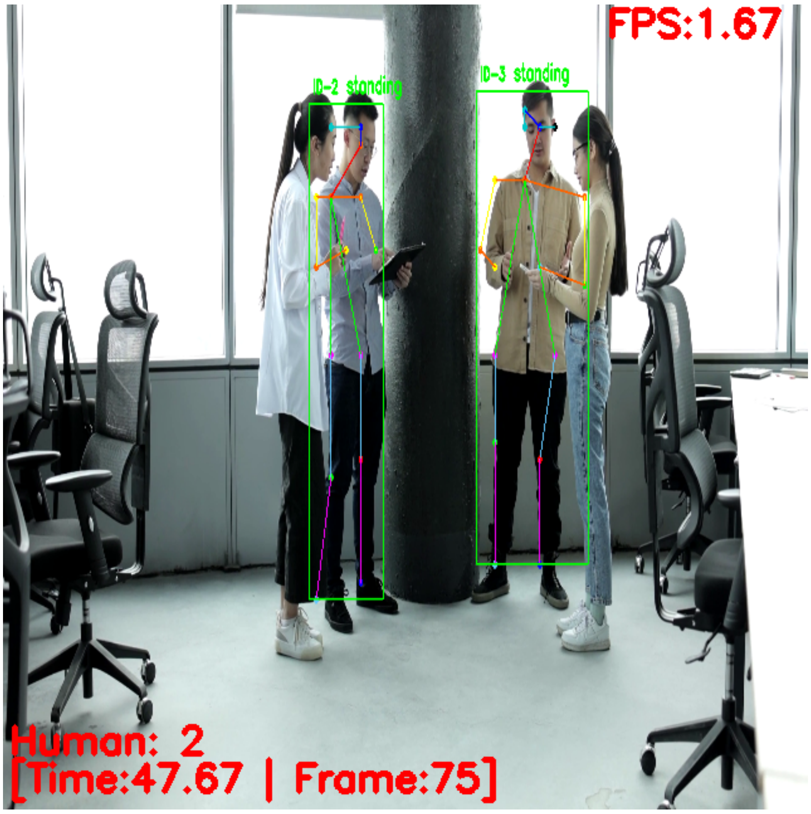


Figure 5.3: Wrong Prediction - Occlusion

5.5 Model Efficiency

In terms of accuracy losses, the model's efficiency can be understood. The confusion matrix is also plotted below:

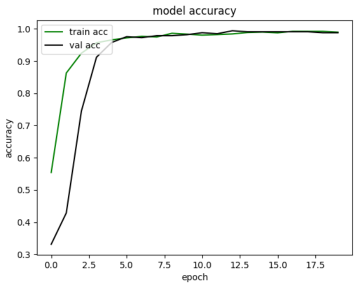


Figure 5.4: Model: Accuracy vs Epoch

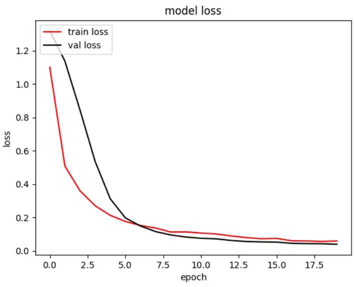


Figure 5.5: Model: Loss vs Epoch

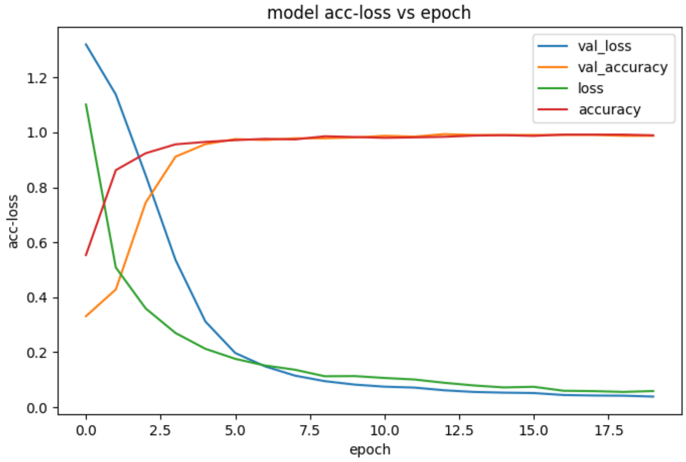


Figure 5.6: Visualization of the model based on traning and validation dataset: Accuracy-Loss vs Epoch.

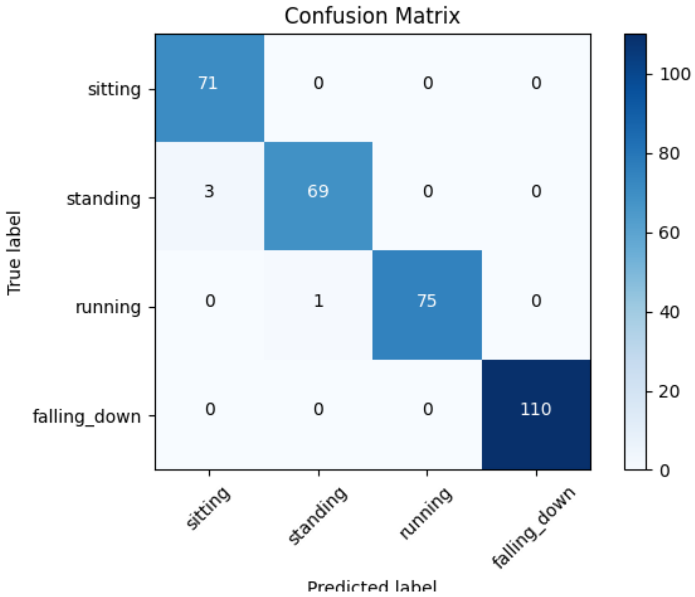


Figure 5.7: The Confusion matrix on the four different types of pose, sitting, standing, running, and falling-down.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The work presented in this paper is mainly focused on investigating an accurate and efficient activities detection for multiple persons based on computer vision techniques. The approach presented here employs the preprocessing model combined with OpenPose and DeepSort to obtain key points and the position information of the human body and then extract four types of features which are sitting, standing, running, falling-down from the preprocessed data. In the detection of these activities, the speed change of the human body part is the main feature, which is tracked by the DeepSort framework. These features are then fed into a CNN model for classification. Finally, the four classification results were combined to obtain the activity detection results.

The approach achieves 97.91% accuracy on the Le2i Fall Detection Dataset and achieves 98.33% accuracy on the UR Fall Detection Dataset. In addition, our method achieves a good balance between sensitivity and specificity. The falling-state features make contribution to identify sudden changes in the shape of the human body. The fallen-state features can identify the state of the human body lying on the ground, and reduce the misjudgment of similar actions, such as bending, down. The fusion of multiple features makes the judgment of falls more reasonable and accurate.

6.2 Future Work

Since, the main purpose of the entire research project is to detect suspicious human activities occurring in public places.

In the future, this approach can be upgraded in the following regions:

- I will improve this method to focus on more complex environments, such as crowded public areas, malls, etc.
- Additionally, the model will be extended to identify other dangerous behaviours that may occur in our daily life.
- Furthermore, I hope to develop a practical application of the activity detection system to integrate with the Internet of things.

Bibliography

[1] Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR abs/1812.08008* (2018).

[2] Casilari, E., Santoyo-Ramón, J. A., and Cano-García, J. M. Umafall: A multisensor dataset for the research on automatic fall detection. *Procedia Computer Science 110* (2017), 32–39.

[3] Gunale, K. G., and Mukherji, P. Fall detection using k-nearest neighbor classification for patient monitoring. In *2015 International Conference on Information Processing (ICIP)* (2015), IEEE, pp. 520–524.

[4] Huang, Z., Liu, Y., Fang, Y., and Horn, B. K. Video-based fall detection for seniors with human pose estimation. In *2018 4th International Conference on Universal Village (UV)* (2018), IEEE, pp. 1–4.

[5] Kwolek, B., and Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer methods and programs in biomedicine 117*, 3 (2014), 489–501.

[6] Lie, W.-N., Le, A. T., and Lin, G.-H. Human fall-down event detection based on 2d skeletons and deep learning approach. In *2018 International Workshop on Advanced Image Technology (IWAIT)* (2018), IEEE, pp. 1–4.

[7] Wojke, N., Bewley, A., and Paulus, D. Simple online and realtime tracking with a deep association metric. *CoRR abs/1703.07402* (2017).

[8] Yu, M., Rhuma, A., Naqvi, S. M., Wang, L., and Chambers, J. A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment. *IEEE transactions on information technology in biomedicine 16*, 6 (2012), 1274–1286.

[9] Yun, Y., Innocenti, C., Nero, G., Gu, I. Y.-H., et al. Fall detection in rgb-d videos for elderly care. In *2015 17th International Conference on E-health Networking, Application & Services (HealthCom)* (2015), IEEE, pp. 422–427.