INTRUSION DETECTION SYSTEM USING OPENPOSE

PROJECT REPORT

Submitted by

DIVYA C P

REG NO : TKM20CSCE05

In partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Under the guidance of

Dr. Dimple A Shajahan

Thangal Kunju Musaliar College of Engineering

Kerala

SEPTEMBER 2022

Thangal Kunju Musaliar College of Engineering
Dept. of Computer Science & Engineering

C E R T I F I C A T E

This is to certify that this report titled ***Intrusion Detection System using OpenPose*** is a bonafide record of the **Project** presented by **DIVYA C P (TKM20CSCE05)**, under the guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **M.Tech in Computer Science & Engineering** in **APJ Abdul Kalam Technological University**.

Coordinator

Dr. Ansamma John
Professor
Computer Science & Engineering
TKMCE

Supervisor & Head of the Department

Dr. Dimple A Shajahan
Professor
Computer Science & Engineering
TKMCE

ACKNOWLEDGEMENT

A successful mini project is a fruitful culmination of efforts by many people, some directly involved and some others indirectly, by providing support and encouragement. Firstly I would like to thank the almighty for giving me the wisdom and grace for making my project a memorable one. I thank him for steering me to the shore of fulfillment under his protective wings.

I express my sincere gratitude to **Dr. T A Shahul Hameed**, Principal, T.K.M College of Engineering for giving me an opportunity to present my mini project. I would like to thank **Dr. Dimple A Shajahan**, Professor and Head of the Department, CSE, TKMCE, for her constant support and encouragement throughout the work.

With a profound sense of gratitude, I would like to express my heartfelt thanks to my guide **Dr. Dimple A Shajahan**, Professor and HOD, CSE, TKMCE, and project coordinator **Dr. Ansamma John**, Professor, CSE, TKMCE for their expert guidance, cooperation and immense encouragement. I also extend my thanks to the entire faculty members and staffs of the Department of Computer Science & Engineering, TKMCE, who has encouraged me throughout this work.

I also express my thanks to my loving parents and friends, for their support and encouragement in the successful completion of this project work.

DIVYA C P

Abstract

Intrusion Detection System(IDS) is a monitoring system that identify unusual activity and send out alerts when it does. It is a core part of site's safety and security strategy. A computer vision task that requires recognizing, associating, and tracking semantic key points is human posture estimation and tracking. Human Action Recognition (HAR) methods are gaining importance, with the emerging advancements in computer vision and pattern recognition. There are various methods present for action recognition and each technique has its advantages and disadvantages. Despite being a lot of research work, action recognition is still a challenging and complex task. The precision of posture estimate has been limited because of the significant processing resources required for semantic keypoint tracking in live video data. Human Pose Estimation is a core problem for understanding people in videos and images. This project presents an effective method for improving existing security systems and to further automate the threat prevention and protection procedure. This method identifies human wall climbing activity, that occur within the video range and detects unlawful or suspicious activities, alert if detected. This work is divided into two parts; multi person pose estimation and action recogonition and alerting using estimated pose. OpenPose library is used for the realtime 2D pose estima- tion and it consists of recognition of 18 body key points and joint locations and a final parsing to get skeleton. These are further used to extract robust motion features, then a Convolution Neural Network (CNN) is used to recognize the activities associated with these features. Different subjects from different camera angles are used to make the approach person-independent. The proposed method shows best result with promising performance, reaching an overall accuracy of 91.6% .

Contents

List of Figures

List of Tables

List of Abbrevations

IDS ..... Intrusion Detection System
HAR ..... Human Activity Recognition
CNN ..... Convolution Neural Network
PAF ..... Part Affinity Field
HPE ..... Human Pose Estimation
NMS ..... Non Max Supression
DNN ..... Deep Neural Network
ReLU ..... Rectified Linear Unit

Chapter 1

Introduction

Intrusion Detection System (IDS) is a monitoring system that identify unusual activity and send out alerts when it does. An IDS's main function is to identify and alert users when someone enters or attempts to enter a guarded area. Any room, a complete building, or a collection of buildings can constitute a secured area. The presence of an IDS at your home or place of business discourages burglars from trying to break in since they have an almost 100 % probability of getting detected. The Intrusion System is the answer to safeguarding your building and the area around it from trespassers, thieves, and other potential intruders. Various application of Intrusion Detection System is shown in Figure 1.1

Figure 1.1: Applications of Intrusion Detection System.

A computer vision task that requires recognizing, associating, and tracking semantic key points is human posture estimation and tracking. Human activity detection and recognition has been a significant subject in the field of computer vision and image processing in the past 30 years. There have been considerable achievements and numerous approaches developed in this field. One of the main challenges in the field of activity recognition is estimating human poses. For the human pose estimation realtime 2D pose estimation method, OpenPose is used and for action recognition Convolution Neural Network (CNN) is used here.

Human Pose Estimation: Human 2D Pose Estimation is a core problem for understanding of people in images and videos.Human estimation has largely focused on finding body parts of individuals. In case of Single Person Pose

Estimation, the problem is simplified by assuming the image has only one person. Because of multiple people in an image Multi Person Pose Estimation is quite more difficult . Multiple people pose inference in images presents a unique set of challenges. Firstly, Unknown number of people can be in each image that can appear at any position or scale. Second, Induced complex spatial interference interactions between people. Third, as the number of people in the image, runtime complexity tends to grow with that making realtime performance a challenge.

There are many works focused on solving this problem over years. To solve the pose estimation for each human, a common approach is to follow a two-step framework which uses a human detector, which makes make the real-time performance a challenge as the running time tended to grow with the number of people in the image. A bottom-up approach is used in this work.

OpenPose is an efficient method presented for multiperson pose estimation with competitive performance on multiple public benchmarks. Bottom-up approach is used where the body parts are detected by the model and a final parsing is used to extract the pose estimation results and its steps are described below:

- Firstly, the entire image is taken as input and it is passed through a baseline network to extract feature maps.
- The feature maps are then processed with multiple stages CNN to generate: 1) a set of Part Confidence Maps and 2) a set of Part Affinity Fields (PAFs). Part Confidence Maps is a set of 2D confidence maps S for body part locations. Each joint location has a map. PAF encodes the degree of association between parts.
- Finally, the Confidence Maps and Part Affinity Fields are finally parsed and post processed to obtain the skeleton pose for each person in the image.

The runtime of bottom-up approach increases relatively slowly with the increasing number of people. Runtime analysis of bottom up approach with number of people is shown in Figure 1.2

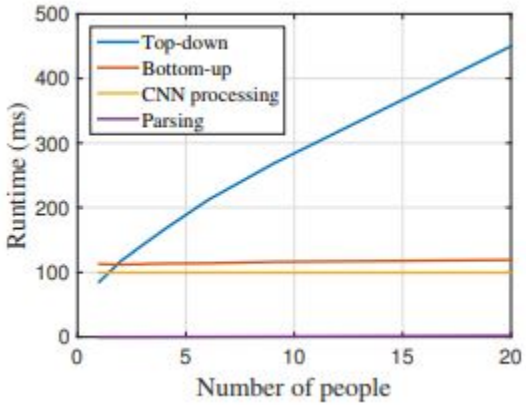

Figure 1.2: Runtime analysis of bottom up approach with number of people.

Action recognition: Human action recognition plays a significant role in video understanding. It has become an active research area in recent years. CNN based action recognition is used here. CNN based system recently demonstrated superior performance in the field of human skeletal identification. To categorise human body motions, it uses a combination of human skeletal identification and deep neural networks (DNN). CNNs the most common deep learning model, categorise and recognise pictures for feature extraction convolution layer is used and for classification pooling layer is used. It is useful for visual data analysis of photos since it can rapidly learn an object's unique feature pattern from an RGB 2D image.

The proposed method involves a database creation using posture determined frames chosen from the wall climbing frames of the input video. CNN model is created and is trained using created database and predicts if someone tries to scale a wall, the system will alert the proper system.

The remainder of this report is organized as follows. Chapter 2 recalls some Related works. The Methodolgy is described in Chapter 3. Chapter 4 presents the Experimental results and evaluation. Finally the Conclusion is provided in Chapter 5.

Chapter 2

Related Works

2.1 Intrusion Detection Systems(IDS)

Intrusion Detection System (IDS) detects unusual behaviour and it sounds an alarm. Based on these notifications, the incident responder can assess the issue and take the appropriate actions to remove the threat.

Tabrizi et.al.[2] proposed a model-based approach for developing intrusion detection systems (IDS) for smart metres that implements IDS on an open source smart metre platform. Demonstrated that IDS has low impact on performance, even when memory is severely limited, and efficiently identifies a variety of known and unexpected attacks. In contrast, current IDSs result in unacceptably high performance overheads. Oludele et.al.[9] proposed a method that shows a system with more than one security device in place tends to prevent unauthorized access and illustrates the implementation of this in an enclosed area whose security level must be kept on the high at all times.

2.2 Action Recognition

Finding out whether a person in an image or video is carrying out a specific action is known as action recognition. A number of acts can be taught to be recognised by AI models. Both people and computer vision models must learn how to recognise and differentiate between various behaviours. Everything involves human behaviour, including home security cameras, movies and television programmes. The need for contextual information in this area of computer vision poses difficulties. Action recognition in computer vision may be applied in the areas of safety and security, healthcare and media..

Yang et.al.[6] proposed a method for arbitrary-view HAR; VS-CNN, a view-guided skeletal CNN technique is used in this method to address the issue of arbitrary-view action recognition. The VS-CNN outperforms other models in experiments, and its dataset contains action samples recorded in 8

fixed perspectives and varying-view sequences that encompass all 360° view angles. A total of 118 people are requested to participate in 40 different action categories, which gives useful but difficult data for the assessment of arbitrary-view recognition. Yan et.al.[28] proposed an approach that investigates methods to enhance activity recognition using human pose skeletons and introduces various methods for producing a reliable representation for training models for activity detection-related challenges.

2.3 Human Pose Estimation (HPE)

HPE is a method for recognising and categorising the joints in the human body. It's a technique for recording a set of coordinates for each joint (arm, head, etc.), which is referred to as a key point that can characterise a person's position. A pair is the relationship between these points. Not all points can pair up since there must be a meaningful connection between them. The initial goal of HPE is to create a skeleton-like representation of the human body, which will subsequently be further processed for task-specific applications. Deep convolutional neural networks outperform all other methods for computer vision tasks, and this is also true in HPE.

Cao et.al.[1] suggested a technique OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields learns to link body parts with persons in the image using a non parametric representation known as Part Affinity Fields (PAFs). No matter how many people are in the image, this bottom-up approach delivers great accuracy and realtime performance. Zhang et.al.[5] proposed a strategy for multi-task pose estimation that deals with human behaviour understanding. It uses multi-task deep learning networks and orientation and occlusion aware multi-person posture estimation. The deep learning model that was employed is based on Mask-RCNN, and its output includes the tasks of mutual occlusion detection, body segmentation, orientation prediction, and human keypoint prediction. The public dataset COCO is used to train the model, which is then supplemented with ground truths for orientation mask and occlusion mask. Wen et.al.[6] proposed Multi-Person Pose Estimation method Using thermal images in this instead of color images with invisible human body parts, ThermalPose that leverages thermal images to accurately track the 2D human pose is used.

Chapter 3

Proposed System

3.1 Overview

Intruder Detection System makes the threat prevention and protection process more automated by combining OpenPose with CNN. This work aims to improve the existing security system and alert whether intrusion by wall climbing is detected. For this action should be recognised properly. HAR is an important research area in the field of computer vision. Main challenge of HAR is human pose estimation. The flowchart of the proposed method is shown in Figure 3.1.

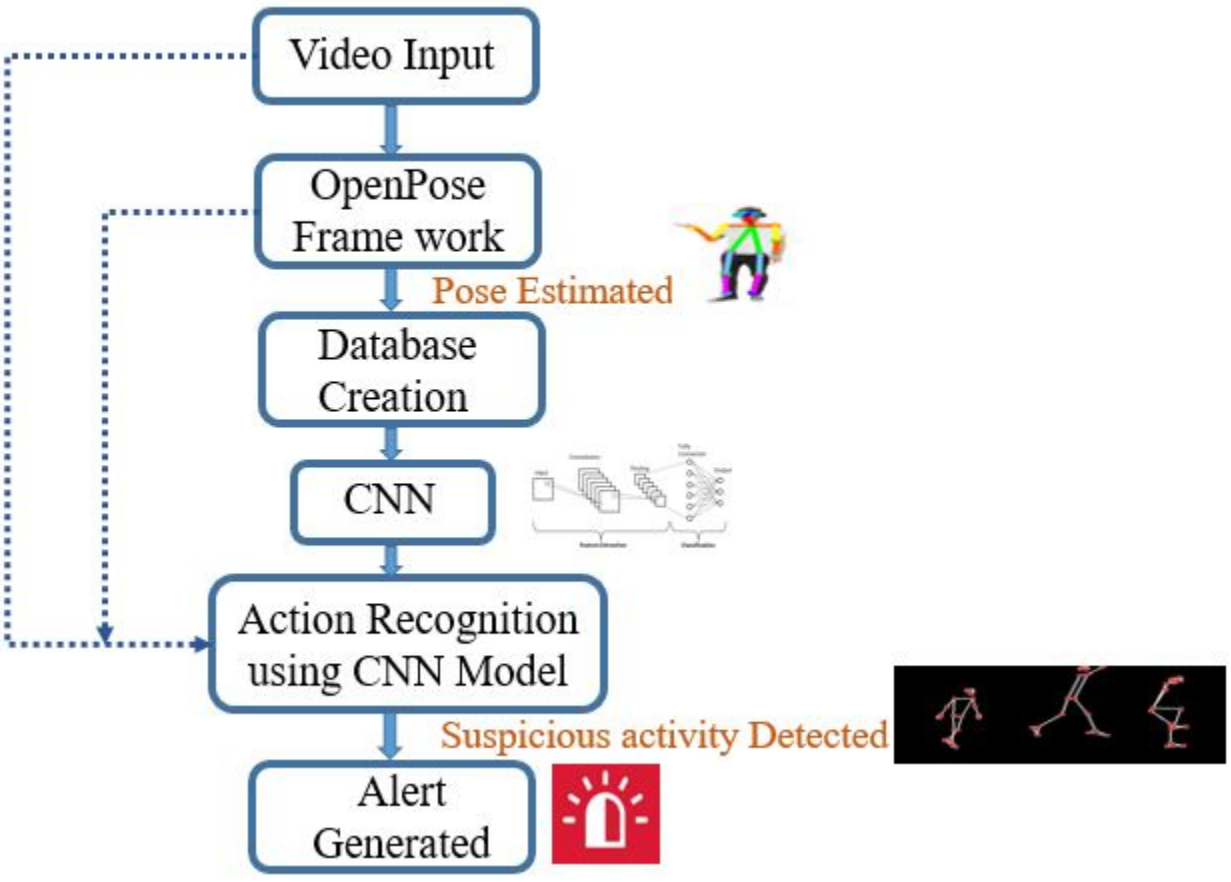

Figure 3.1: Flowchart of proposed IDS using OpenPose

The proposed method is divided into 3 phases; Database Creation, CNN model Creation, implementation of action recognition CNN model with OpenPose and alerting.

A previously captured wall climbing video is fed as the input to OpenPose framework. OpenPose gives 2D realtime pose estimation, then a database is created from the estimated poses of input video frames. CNN is trained with the created database to make the model for recognition. For the action recognition, CNN model was given with input video and OpenPose framework to predict the activity. Finally If any unusual activity, wall climbing of intruder detected alert is produced.

3.2 Pose Estimation Using OpenPose

OpenPose is a fast method for estimating the realtime poses of multiple people that achieves the highest accuracy on a variety of open benchmarks. The architecture of the OpenPose method is shown in Figure 3.2.

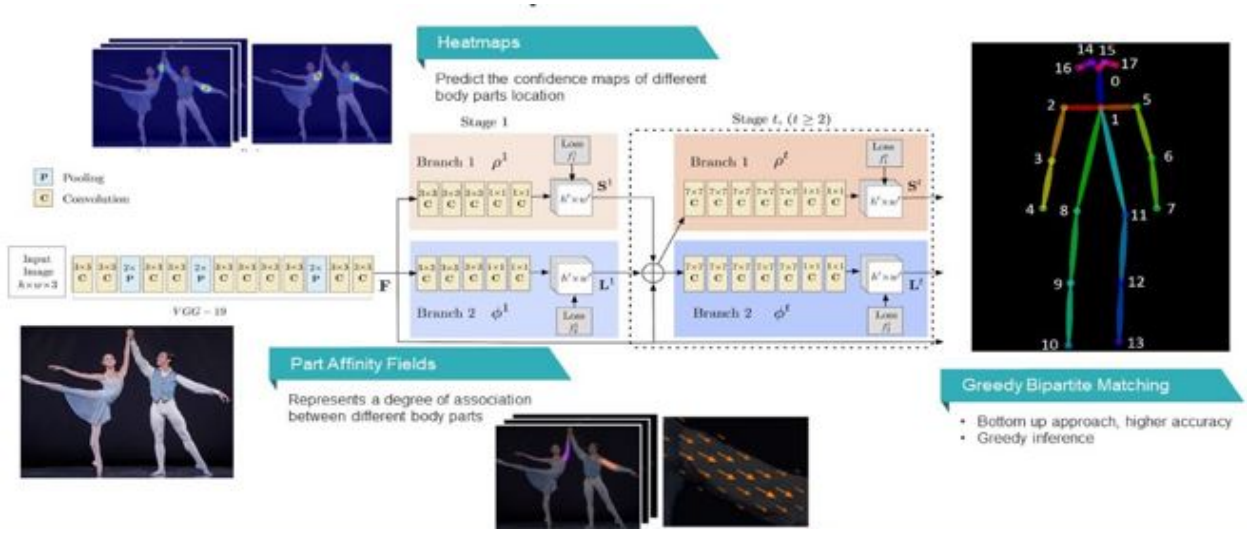

Figure 3.2: OpenPose architecture.

It introduces Part Affinity Fields (PAF), the first bottom-up depiction of association scores. A group of 2D vector fields called PAFs that encode the position and limb positioning inside the picture domain. The inferring of bottom-up representations of detection and association at the same time en-

codes global context effectively enough to enable a greedy parse to attain superior outcomes for a small fraction of the computing cost. This method presented the first real-time pose recognition system for several 2D characters and publicly shared the source for full reproducibility. The Overall pipeline of OpenPose architecture is shown in Figure 3.3.

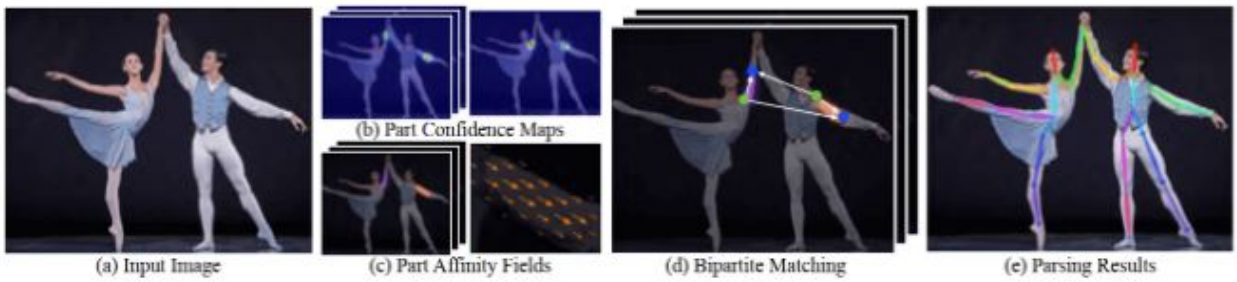

Figure 3.3: Overall pipeline of OpenPose architecture.

Steps involved in OpenPose are Feature extraction, Confidence Map generation, Part Affinity Fields(PAFs) generation, Post processing to get skeleton.

An RGB image of size $h \times w \times 3$ is fed into VGG-19 for the features extraction as shown in Figure 3.2 . Extracted features are then passed into two split networks of multi-stage CNN. These two branches will predict 2 different things. First branch predict set of 18 heat maps. Each will give which part we are trying to estimate. Here a total of 18 keypoints are estimating so 18 heat maps. Second branch gives different output which represents degree of association between body parts; ie.,it gives the info about which of the body parts can be joined together to give a pair. Each stage CNN refine whatever the output come from previous stage.

Finally, post processed to find which of the parts and pairs are combined to create human skeleton. Post processing involves; applying Non Max Supression (NMS) to get the confidence of each body parts. For each part, not a single pixel value is getting. Here, the PAF comes into role and along with line integral all possible connections are found out based on heighest weightage value, sort it from max to min. Merge all the pairs that go with each human. For this same number of humans have whatever connections obtained, each connection belong to different person both pairs having common

vertices is considered as single joint for single human skeleton; represented as

$$if \;\; H_1 \cap H_2 \neq \phi$$

then

$$H_1 = H_1 \cup H_2$$

$$delete(H_2)$$

3.2.1 Confidence Map

A confidence map is a two-dimensional depiction of the idea that a certain body component may be located. On a single map, a single body part will be depicted. Sample of Confidence map is shown in Figure 3.4

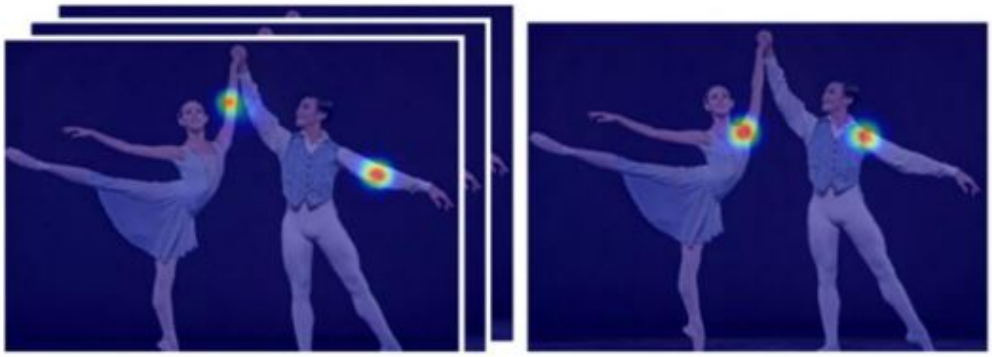

Figure 3.4: Confidence map of left hand elbow and left hand shoulder.

The first branch of feed forward network shown in Figure 3.2 predicts a set of 2D confidence maps of body part locations S and a set of 2D vector fields of part affinity fields (PAFs) L, which encode the degree of association between parts. The set of 2D confidence maps $\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k)$ has K confidence maps, one per part, where $\mathbf{S}_k \in R^{w \times h}$, $K \in \{1 \ldots k\}$. The set $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_C)$ has C vector fields, where $\mathbf{L}_c \in R^{w \times h \times 2}$, $c \in \{1 \ldots C\}$ i.e, one per limb, which refers to part pairs as limbs, but some of the pairs are not human limbs Each image location in Lc encodes the 2D vector. Finally, the confidence maps and the PAFs are parsed by greedy inference to give the 2D keypoints for all people present in the input image.

3.2.2 Part Affinity Fields (PAF)

PAF is required, notably in multi-person posture detection, where it must map the correct body parts to its body. Because many people have several heads, hands, shoulders, and so on. When they are clustered together, it may be difficult to tell them apart. PAF links many parts of the body that belong to the same person. A greater PAF association between bodily components implies that they belong to the same person. PAF is shown in Figure 3.5

Figure 3.5: Part Affinity Fields (PAFs).

3.2.3 Multi-Person Parsing Using PAFs

Part locations are extracted out from the heatmap, after that right connections have to be find out. PAF comes into role to find right connections. Then merging is done to transform detected connections into final skeletons. Samples of estimated realtime 2D Pose is shown in Figure 3.5

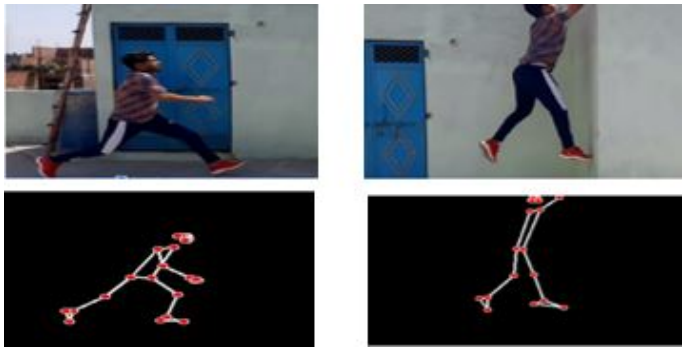

Figure 3.6: Samples of estimated pose Using OpenPose.

The graph below indicates that, in contrast to top-down techniques such as Mask-RCNN and AlphaPose, OpenPose has essentially little influence on the

number of persons present in the image. Comparison of Openpose with other techniques is shown in Figure 3.7

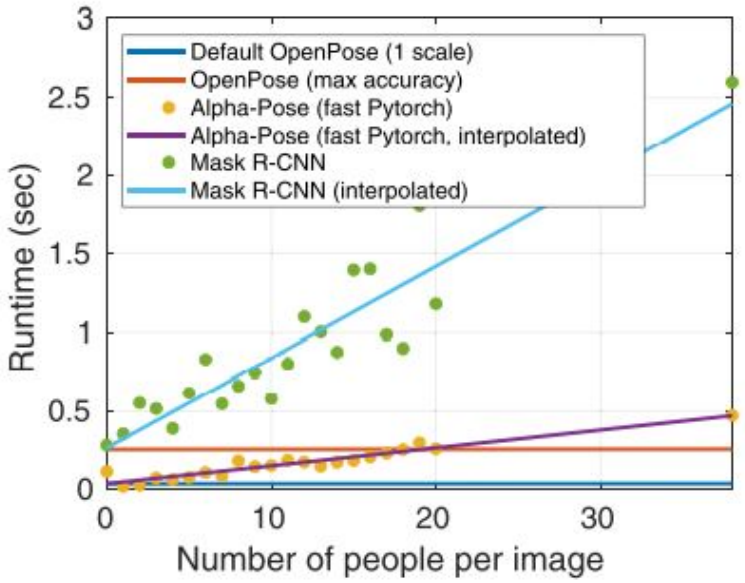

Figure 3.7: Comparison of OpenPose with other techniques.

3.3 Action Recognition Using CNN Model

CNN based system recently demonstrated superior performance in the field of human skeletal identification. To categorise human body motions, it uses a combination of human skeletal identification and Deep Neural Networks (DNN). CNNs, the most common deep learning model, categorise and recognise pictures using a convolution layer for feature extraction and a pooling layer for classification. It is useful for visual data analysis of photos since it can rapidly learn an object's unique feature pattern from an RGB 2D image.

CNN-based model has three kinds of layers: 1) An input layer; 2) hidden layers whose values are derived from previous layers ; 3) and output layer whose values are derived from the last hidden layer.

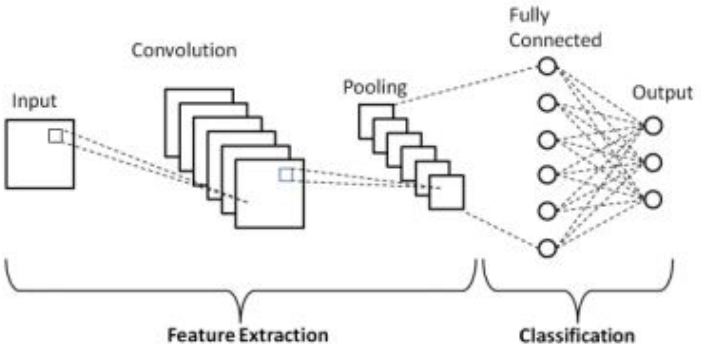

Figure 3.8: Layers of CNN.

Figure 3.8 shows the layers of CNN. The feature extraction inputs of each neuron are directed to the convolutional layer (C layer). It is connected to the local receptive field of the previous layer and extracts the local function just once. A positional relationship was formed between the local feature and the other features. The pooling layer (S layer) is a supplemental blur filter. Extraction of features that alters the spatial resolution of buried layers.

The number of layers that may be used in each layer to extract multiscale information from images of human activities. To incorporate nonlinearity, the Rectified Linear Unit (ReLU) beat the prior sigmoid function due to several advantages, including the computation of partial derivatives of ReLU is easy, training time is short.

Flow of Pose estimation and HAR process is shown in Figure 3.9

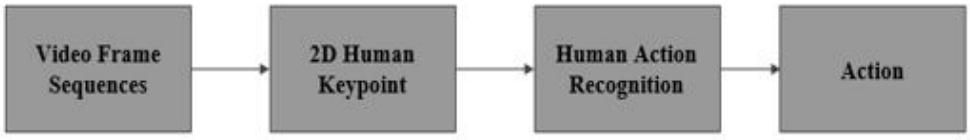

Figure 3.9: Pose estimation and HAR process flow.

The wall climbing frames are chosen from the posture determined frames and

used to train the system using the CNN model. If someone tries to scale a wall, the system will alert the proper system.

3.4 Evaluation

Proposed method is evaluated using confusion matrix. Accuracy and F1-Score is measured and compared with other models.

Accuracy represents the number of correctly classified data instances over the total number of data instances and it is defined as follows:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \tag{3.1}$$

where TN, TP, FP and FN are True Negative, True Positive, False Positive and False Negative respectively.

F1-Score is a metric which takes into account both precision and recall. F1-Score is defined as follows:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.2}$$

where

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.3}$$

and

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.4}$$

Chapter 4

Experimental Results & Discussion

Dataset was created using video frames. To test the performance of this approach, which comprises various walking and climbing wall positions collected from various films. Some of the specific actions from the created database is shown in Figure 4.1

Figure 4.1: Sample of created database .

Using the OpenPose library, which is ideal for human body pose estimation in real time for single and multi-person video analysis, real time pose from

the input is estimated. Real time estimated pose inorder to identify action is shown in Figure 4.2

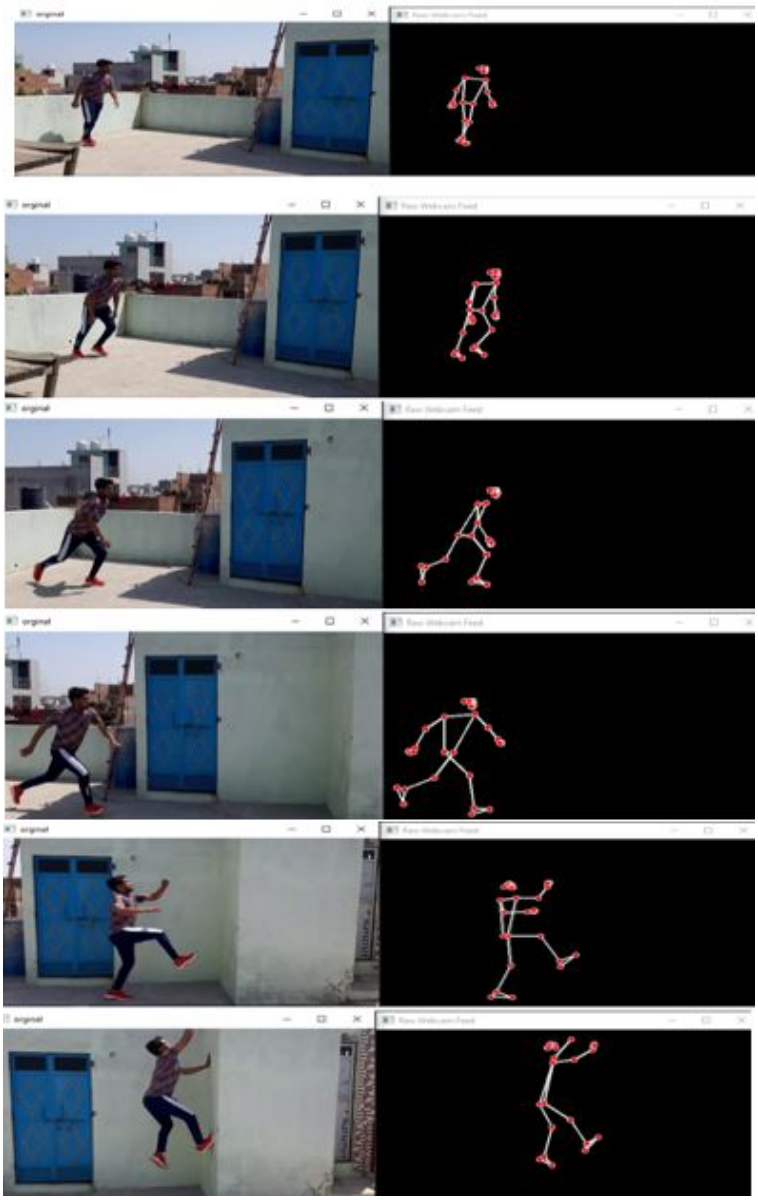

Figure 4.2: Samples of real time estimated pose from input to identify action.

When compared to other models for posture estimation and action identification, this model performs the best. Accuracy and F1-Score of OpenPose model is evaluated using Equation (3.1) Equation (3.2). Accuracy and F1-Score Comparison of models is shown in Table 4.1

Table 4.1: Accuracy and F1-Score Comparison of models

Model	**Accuracy (%)**	**F1-Score (%)**
CNN with OpenPose	93.6	93
LSTM	85.6	85.4
CNN-LSTM	88.68	88.98
SVMs	84.07	83.76

The confusion matrix for the datasets is shown in Figure 4.3

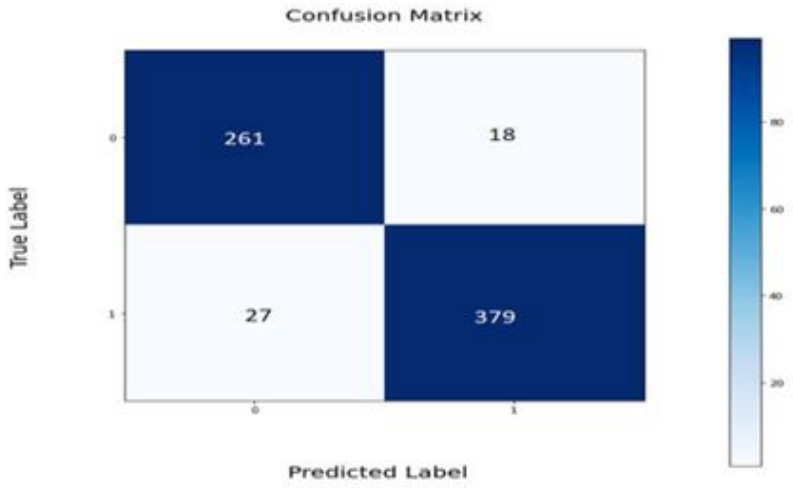

Figure 4.3: Confusion matrix for the datasets.

The confusion matrix for the datasets for true and predicted labels is depicted in Figure 4.3. In this case, 0 signifies no wall climbing and 1 represents wall climbing. The matrix shows that the model successfully predicted wall climbing 261 times and incorrectly predicted wall climbing 27 times. Similarly, the model accurately predicted no wall climbing 379 times and incorrectly predicted no wall climbing 18 times.

Training and testing accuracy of the proposed method shows that accuracy doesn't fall below 93% and is shown in Figure 4.4.

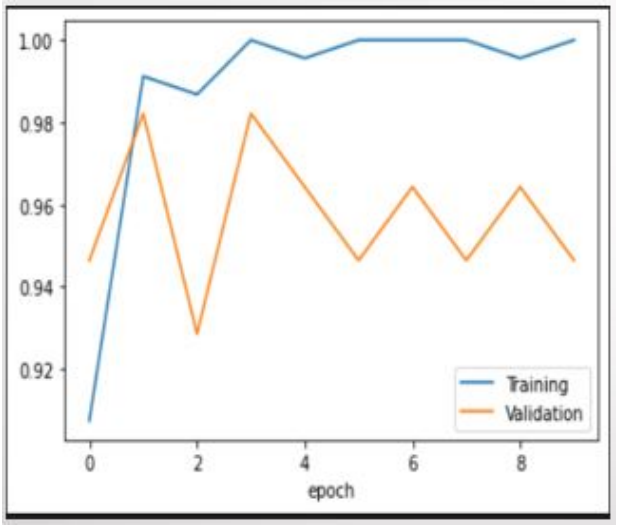

Figure 4.4: Training and Testing accuracy

The wall climbing frames are chosen from the posture determined frames and used to train the system using the CNN model. If someone tries to scale a wall, the system will alert the proper system.

Detected wall climbing action is also displayed when detected, it is shown in Figure 4.5

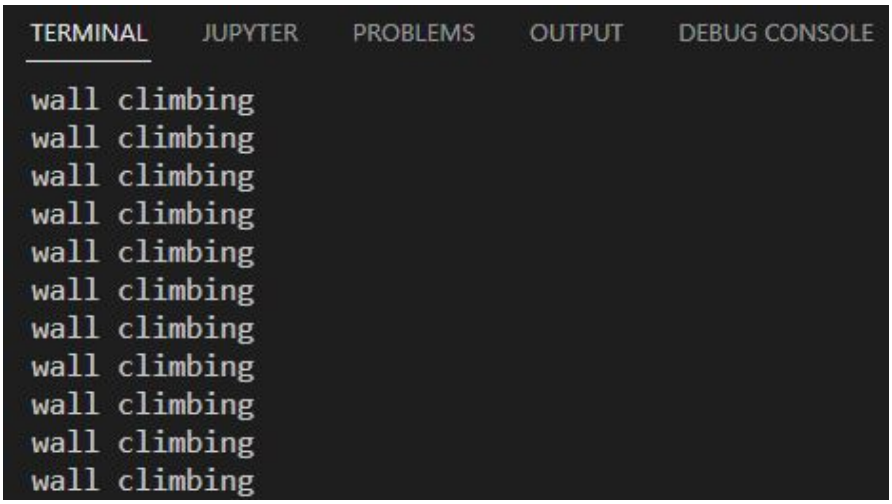

Figure 4.5: Detected wall climbing action.

If any suspicious activity detected then the system will alert the correspond-

ings. Output Screenshot of suspicious activity detection is shown in Figure 4.6

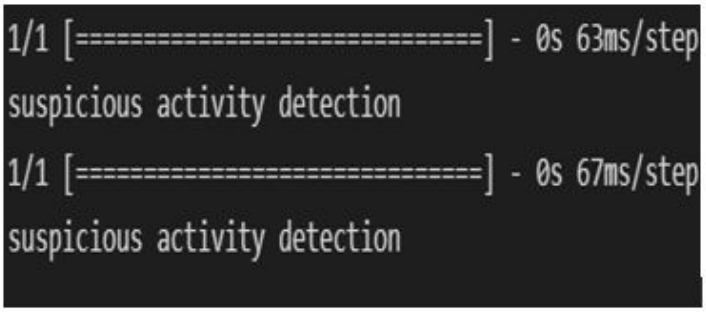

Figure 4.6: Output Screenshot of detected suspicious activity.

Chapter 5

Conclusion & Future Works

Deep convolutional neural networks have been successfully used to learn a variety of computer vision tasks. In this section, it will be clarifying the fundamental knowledge and principles of CNN and the deep learning technique for human posture assessment. The detection of human activity is an important research area in pattern recognition and ubiquitous computing. The model initially predicts the 2D locations of body joints from raw RGB frames. These places are used to forecast the activities in the video. This is one of the most recent improvements in deep learning techniques to video-based activity identification. Deep learning increases performance over standard pattern recognition algorithms by lowering reliance on human-generated feature extractions and automatically learning high-level representations of video-based data. Pose estimation is an intriguing component of computer vision that may be utilized in a variety of industries, including technology, healthcare, and business. It is utilised for security and surveillance systems, in addition to modeling human personalities using Deep Neural Networks that learn numerous critical elements. This method uses CNN along with OpenPose. Performance accuracy of this method obtained is 91.6%.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields", IEEE Transactions on Pattern Analysis and Machine Intelligences, vol. 43, 2021.

[2] F. M. Tabrizi and K. Pattabiraman, "A Model-Based Intrusion Detection System for Smart Meters," 2014 IEEE 15th International Symposium on High-Assurance Systems Engineering, 2014, pp. 17-24, doi: 10.1109/HASE.2014.12.

[3] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4929–4937.

[4] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multiperson pose estimation model", IEEE Trans. on Comput. Vis., 2016, pp. 34–50.

[5] H. Zhang, Y. Gu and S. Kamijo, "Orientation and Occlusion Aware Multi-Person Pose Estimation using Multi-Task Deep Learning Network," 2019 IEEE International Conference on Consumer Electronics (ICCE), 2019, pp. 1-5, doi: 10.1109/ICCE.2019.8662041.

[6] C. Chen, C. -J. Wang, C. -K. Wen and S. -J. Tzou, "Multi-Person Pose Estimation Using Thermal Images," in IEEE Access, vol. 8, pp. 174964-174971, 2020, doi: 10.1109/ACCESS.2020.3025413.

[7] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Strike a Pose: Tracking people by finding stylized poses", in Proc. IEEE Trans. Comput. Vis. Pattern Recognit., 2005, pp. 271–278.

[8] Y. Ji, Y. Yang, F. Shen, H. T. Shen and W. -S. Zheng, "Arbitrary-View Human Action Recognition: A Varying-View RGB-D Action Dataset," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 1, pp. 289-300, Jan. 2021, doi: 10.1109/TCSVT.2020.2975845.

[9] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition", Int. J. Comput. Vis., vol. 61, pp. 55–79, 2005.

[10] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts", IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[11] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770-778, 2016.

[12] Awodele Oludele, Ogunnusi Ayodele, Omole Oladele and Seton Olurotimi"Design of an Automated Intrusion Detection System incorporating an Alarm" Journal of computing, Volume 1, Issue 1, Decmber 2009, ISSN: 2151-9617

[13] Z. Cao, T. Simon, S.-E. Wei and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1302-1310, 2017.

[14] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4645–4653.

[15] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines", in Proc. Eur. Conf. Comput. Vis., 2014, pp. 33–47.

[16] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3D deformation model for tracking faces, hands, and bodies", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8320–8329.

[17] S. . Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Realtime 3D human pose estimation with a single RGB camera", ACM Trans. Graph., vol. 36, 2017, Art. no. 44.

[18] J. . Lan and D. P. Huttenlocher, "Beyond trees: Common-factor models for 2D human pose recovery", in Proc. IEEE Int. Conf. Comput. Vis., 2005, pp. 470–477.

[19] M. Eichner and V. Ferrari, "We are family: Joint pose estimation of multiple persons", in Proc. Eur. Conf. Comput. Vis., 2010, pp. 228–242.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[21] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "Using k-poselets for detecting people and localizing their keypoints", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 3582–3589.

[22] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations", in Proc. Eur. Conf. Comput. Vis. Workshop, 2016, pp. 627–642.

[23] G. Moon, J. Chang, and K. M. Lee, "Posefix: Model-agnostic general human pose refinement network", CVPR, 2019.

[24] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos", in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1913–1921.

[25] J. J. Tompson, A. Jain, Y. LeCun and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation", Proc. Advances Neural Inf. Process. Syst., pp. 1799-1807, 2014..

[26] S.-E. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh, "Convolutional pose machines", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4724-4732, 2016.

[27] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression", in Proc. Eur. Conf. Comput. Vis., 2016, pp. 717–732.

[28] M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3686-3693, 2014.

[29] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4645–4653.

[30] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, "Learning to refine human pose estimation", in Proc. Comput. Vis. Pattern Recognit. Workshop, 2018, pp. 205–214.

[31] B. Raj N., A. Subramanian, K. Ravichandran and N. Venkateswaran, "Exploring Techniques to Improve Activity Recognition using Human Pose Skeletons," 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW), 2020, pp. 165-172, doi: 10.1109/WACVW50321.2020.9096918.