IoT-based Human Activity Recognition using Deep learning

Salman Ahmad Siddiqui[1,*], Anwar Ahmad[2], Ankur Varshney[3]

[1]Department of ECE, Jamia Millia Islamia, New Delhi
[2]Department of ECE, Jamia Millia Islamia, New Delhi
[3]Amdocs Development Center India LLP, Gurgaon, Haryana

Abstract

Artificial intelligence and the Internet of things (IoT) are the fastest and latest growing technologies that can handle a huge amount of data in computing services. This paper presents a smart human activity recognition system based on IoT that can be used for surveillance purposes working as IoT-based armour. Pose estimation model viz. MoveNet has been employed to extract the anatomical key points from RGB video frames. Different subjects from different camera angles were employed to make the approach person-independent. Diverse Machine learning models such as Decision tree, support vector machines, XGboost, and random forest classifiers were employed using extracted keypoints for training the model for estimating human activity during posture estimation monitoring. SMS are sent to the designated person with the raising of buzzer alarm in case of anomalous behaviour detection.

Keywords: Artificial intelligence, Internet of things, MoveNet, Pose estimation, Machine learning.

Received on 09 September 2022, accepted on 05 April 2023, published on 20 April 2023

doi: 10.4108/eetcasa.v9i1.2682

1. Introduction

In today's world, human activity recognition (HAR) has become one of the vital areas of deep learning and has grabbed much attention from the research community. Deep learning based new methods are replacing traditional computing methods and are providing excellent solutions for medical, finance, speech recognition, bioinformatics and other fields.

Human activity in video analysis has become topic of interest and is being employed in different domain applications such as surveillance, physical activity, sports, health care etc.

Based on literary sources, HAR is carried out to distinguish and analyse motions based on wearable sensors or some sensing modalities. These wearable sensors are attached to human body for long time, making movement difficult and annoying many times. Sensors such as kinects and RGB have gained much popularity in HAR applications. The light-environment problem is solved with Kinect-based action recognition, which reliably records the skeletal joints during activity.

Although applicable for a variety of applications, traditional approaches to activity recognition are frequently slow and lack stability and performance in complicated environments. This paper proposes deep learning approach using multipurpose MoveNet model and machine learning algorithms to recognise activity. Further object detection using ssd mobilenet has been incorporated with this to give human activity recognition a better picture.

MoveNet is a high-speed, high-accuracy model that detects 17 body keypoints. The Lightning and Thunder variations of the model are available on Tensorflow Hub. Lightning version has been employed in this work.

The present study provides a novel mechanism to detect the human pose on live video streaming capable on different angels and extracting the anatomical keypoints based on different human activities with multiple subjects (humans) for each case. These activity based keypoints are used to train machine learning (ML) algorithms such as random

[*] Corresponding author. Email:salman007.rec@gmail,com

forest, xgboost, Support vector machine (SVM) and decision tree for classifying the type of activity detected. Then alerts using sim800l GSM module and buzzer working with raspberry pi are sent to apprise the officials about the conditions. As a result, the IoT is defined as the use of these interlinked and associated sensors in this study that can gather and transfer information over the network without any supervision. This internetworking enables advanced IoT applications such as smart environments, surveillance, and so on.

The IoT based armour will identify and certify crimes in real-time, using Artificial Intelligence -to dispatch crime data to protective services and police units facilitating instant action while saving resources. This system when installed near homes or office vicinity constitute smart environment.

After occurrence of an incident, authorities typically visit the scene, personally grab the content from the camera, and then proceed to select pertinent footage by watching the entire video or processing it through sophisticated video analytics algorithms. They frequently take a reactive strategy, relying on witness testimony or closed-circuit television (CCTV) footage after a crime has occurred.

Thus with many sensors in place and real time AI based smart video surveillance this system provides instance identification and reporting of human activity making it a smart system.

The paper is divided into sections that are rationally organised. Section II talks about the related works. Section III delves into the system's technique in depth, with a focus on the many machine learning models employed in the system. Section IV offers the findings as well as an in-depth study of the resolution of the most efficient algorithm and the resulting results. Section V discusses the challenges of the developed system with VI presenting the discussion of the proposed work. Finally, the conclusion is present in Section VII.

2. Related Work

Numerous research has been carried out in this field each having benefits and drawbacks. This section brief out different works carried out in this area of interest.

In [1], the paper presents a smart video surveillance system executing AI algorithms; the employed AI application allows detecting people in the surveillance area using MobileNet-SSD architecture and kalman filter banks and achieves a precision of 81.43% and a recall of 80.6% in the EPFL-corridor dataset and a portable video surveillance system with AI CNN processing at the edge is also proposed.

The paper [2] develops a hybrid model by incorporating CNN and LSTM for activity recognition with dataset generated from 20 participants using the Kinect V2 sensor and contains 12 different classes for activities achieving an accuracy of 90.89% with this technique.

The paper [3] proposes three deep learning architectures, to perform a joint detection and pose estimation, by decoupling the two tasks using PASCAL3D+ and ObjectNet3D datasets. In [4], a framework for spatial regression using mixture density networks has been designed for both object detection and human pose estimation. For object detection the mixture model learns to deal with object scale variation through different components and in human pose estimation, a mixture model divides the data based on viewpoint and uncertainty. The paper [5] analyses current video datasets for violence detection and proposes the RWF-2000 database with 2,000 videos taken via cameras in real-world scenarios and present a new method that utilizes the benefits of 3D-CNNs and optical flow, namely Flow Gated Network. The proposed approach attains an accuracy of 87.25%. The paper [6] proposes a method to merge the robustness of CNNs with a fine-resolution instance-based 3D pose estimation where the models is trained with fully annotated synthetic training data of the 3D models of the object. Personalized machine learning and deep learning strategies were studied in [7], and their performance was compared to typical deep learning methods. Motion Sense, MobiAct, and UniMiB SHAR were three of the most commonly utilised datasets in this study that contain physical information about the participants. Via amalgamating RPN and Fast R-CNN into a single network by sharing their convolutional features—by using the trendy terminology of neural networks with 'attention' mechanisms—the study in [8] suggests a Region Proposal Network (RPN) that apportionment image convolutional attributes with the detection network, allowing for near-cost-free region proposals, and amalgamates RPN and Fast R-CNN into a lone network by sharing their convolutional features. The paper [9] propose a methodology for an automatic real-time video-based surveillance system which can simultaneously perform the tracking, semantic scene learning, and abnormality detection in an academic environment and recognize and categories actions into two categories: normal and aberrant based on SVM. The study in [10] employs technique for human activity recognition that extracts anatomical important points from RGB photos using the open source library Open- Pose. These key points are then used to generate robust motion features based on their movements in consecutive frames.' Then, on a publicly available activity data set, a Recurrent Neural Network (RNN) containing Long Short-term Memory cells (LSTM) is used to recognise the activities related with these attributes, with an overall accuracy of 92.4 percent. In [11], paper present a system for dynamic robot grasping that conducts real-time object identification and position estimation. While in [12], the field of application, acquisition technology, computer vision techniques, and classification strategies of vision-based pedestrian detection systems are all examined. The paper in [13] proposes a novel convolutional network design for the job of human posture estimation, demonstrating how repetitive bottom-up, top-down processing combined with intermediate supervision is important to the network's performance, and referring to the architecture as a "stacked hourglass" network. In this [16], current multi-view techniques for human 3D posture estimation and activity identification are reviewed and compared. We go over the requirements for the application domain of human pose estimation and activity recognition. Employing two publicly accessible datasets, they compare the most effective techniques for multi-view human action recognition. Using only acceleration and gyroscope data, paper [17] provides a unique optimal activity graph generation algorithm that incorporates a deep learning framework for automatic and accurate HAR with many participants. The multisensory integration process is presented in the activity graph creation model, which uses three-step sorting algorithms to create the best activity graphs with aligned neighbouring signals in both breadth and height. In order to automatically extract distinguishing features from the graphs for HAR, they suggest using a deep convolutional neural network. In order to recognise human action in movies, provide a collection of kinematic properties that are obtained from the optical flow. Divergence, vorticity, symmetric and antisymmetric flow fields, second and third principal invariants of flow gradient and rate of strain, and third principal invariant of rate of rotation tensor are included in the collection of kinematic features. When each kinematic characteristic is calculated from the optical flow of a set of photos, a spatiotemporal pattern is created.

Thus by analyzing different research works, then by deploying the best method for distinct purposes, the manifested framework is planned to provide ways that take video stream for human estimation and then most efficient machine learning model trained on the extracted key point is called for activity prediction and if any abnormal activity such as fighting, person laying on road etc is detected, then sim800l gsm module sent SMS alerts having information about distance from the camera and type of activity with buzzer alarm going up. In case of unscrupulous activity, an object detection model trained using ssd_mobileNet_ for detecting presence of knife or gun is called to further estimate the seriousness of the scene.

Multiple cameras at different locations can be installed as the system is designed in such a way to work from different points. Pi camera with raspberry pi and sensors (buzzer and sim800l) is placed at edge node which results in decrease in latency and boost in network speed. Fog can help network edge devices and end-user devices (such as vehicles, drones etc) form local networks, offering temporary security credentials to these local devices to assist them establish trustworthy communications, and operate as local application servers and data storage servers for edge networks. Server for computing deep learning task is positioned at fog node. Thus fog node reduces the processing burden of edge node and also constitute in optimum network communication speed due to its nearness to the edge node.

3 Methodology

The study put forwards an IoT based smart human activity recognition for surveillance purpose with pose estimation combine with activity recognition based on machine learning models as discussed in below subsections. Further object detection model is called each time with pose estimated video frame to detect for possible weapons to attack and then alerts are sent to take appropriate steps as shown in figure 1 below.

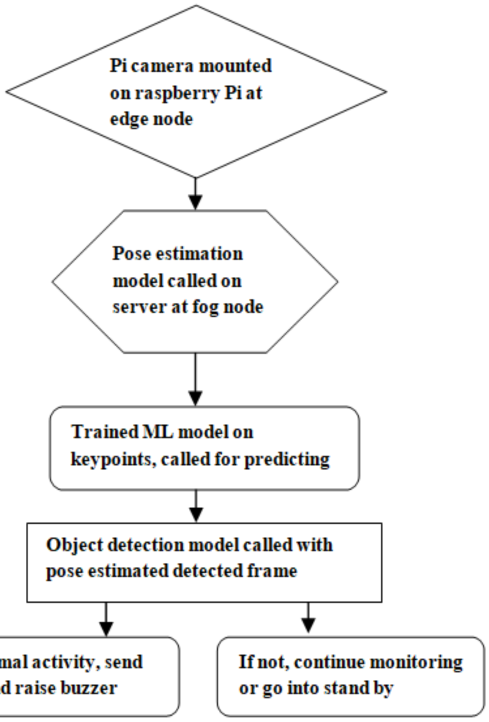

Pi camera mounted on raspberry Pi at edge node
Pose estimation model called on server at fog node
Trained ML model on keypoints, called for predicting
Object detection model called with pose estimated detected frame
If abnormal activity, send alerts and raise buzzer
If not, continue monitoring or go into stand by

Figure 1. System overview

3.1 Dataset Description

For training the models with extracted keypoints, dataset was generated from multiple subjects with 5 classes namely running, walking, fighting, laying, Jumping. Multiple videos with different number of persons (one, two, four and so on) and diverse durations like 3 sec, 6 sec, 8 sec and 10 sec etc were taken for keypoints extractions and the label being the activity. Since we have varied type of classes, different video sequences were collected from different dataset. Video sequences from dataset [14-15] were employed for jumping, running, walking classes. Many videos were taken from YouTube depending on the class namely fighting. To make dataset more diverse in terms of videos sequence many real time videos were generated for the experiment for classes such as running, walking. Thus dataset has different length sequences with different number of frames. The videos were having different resolution, so they were scaled down to a common resolution of 640* 480 with 25fps. The extracted keypoints were stored in CSV file for training the models. The dataset (csv file) does not contain any null values making up to 4157 rows of dataset. Section 3.3 lists the machine learning models. They were assessed based on certain parameters such as macro average f1 score, accuracy, precision and recall (sensitivity) metrics (Table 1). The equations for them are represented by (1)-(4), correspondingly.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} * 100 \qquad (1)$$

$$\text{Precision}= \frac{TP}{TP+FP} \qquad (2)$$

$$\text{F1 Score}= \frac{2*sensitivity\ *precision}{sensitivity\ +precision} \qquad (3)$$

$$\text{Sensitivity} =\frac{TP}{TP+FN} \qquad (4)$$

3.2 Pose estimation

Pose estimation is a computer vision method that tracks and forecasts the pose of a person as shown in figure 2. This is accomplished by estimating the keypoints. For human these keypoints are joints such as elbow, knee, ankle etc. One can follow an object or person in real-world space at an extraordinarily detailed level using pose estimation. Object detection is a task that locates objects inside an image. However, this localization is usually coarse, consisting of a bounding box that encompasses the object. Pose estimate goes much further, estimating the exact location of the object's keypoints.

There are two approaches for pose estimation: bottom-up and top-down. In bottom-up approach, first every instance of keypoints are detected by model and then these are grouped and assembled into skeletons while in top-down, first object detector is employed to draw boundary boxes and then keypoints are determined in each boxes. There are too many specific architectures for pose estimations like mediapipe, open pose, posenet each having their own set of advantages and limitations. MoveNet (lightning version) have been employed for this study.

A Bottom-up estimation model, movenet consist of two parts: feature extractor and a set of prediction heads. The lightning variant runs at 30FPS and can track up to 6 people simultaneously with high-quality performance as compared to mediapipe which track single person.

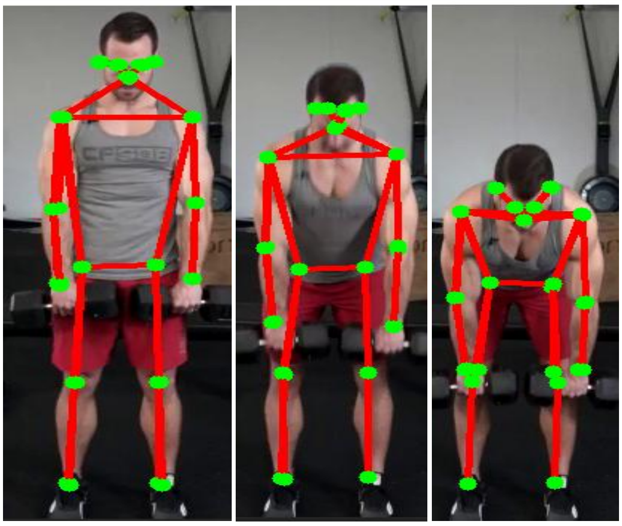


Figure 2. Single Person MoveNet Pose Estimation

3.3 Machine Learning Models

ML models have been employed for training and comparison for estimating the best model based on the above comparison matrix. Models employed are random forest, xgboost, Support vector machine (SVM) and decision tree.

Support Vector Machine (SVM) is a machine learning algorithm that may be used to solve both regression and classification issues. To categorize n-dimensional space into classes, the SVM technique relies on the premise of finding the most optimal line or decision boundary.

A random forest classifier is made up of a number of decision trees that are trained on distinct subsets of the dataset. The average is then used to improve the dataset's forecasting accuracy. Instead of depending on a single decision tree, the final result is anticipated based on the majority of votes.

Another prominent ensemble learning approach with gradient boosting is XGBoost, or Extreme gradient boosting, which enhances the speed and performance of tree-based (sequential decision trees) machine learning algorithms.

Decision Trees are a type of Supervised Machine Learning in which data is continually separated based on a parameter to generate decisions.

3.3.1 Pose classification

The dataset prepared by the extracted keypoints was used to train the ML models by dividing the dataset into train and test halves in 75:25 ratios and the results are tabulated in table 1.

Random forest and decision tree was trained using hyperparameter tuning for optimizing the result. Varied combination were explored and tested. In the case of the decision tree, the entropy criterion was used, and 200 trees with the entropy criterion were chosen in the case of the random forest. For the classification job, the best performing model (see table 1) is finally used.

3.3.2 Employed Object detection model

Since the task also focuses on finding the presence of dangerous weapons in the hands of a person, an object detection model was trained using ssd_mobilenet_v2_fpnlite_320x320_coco17_tpu-8. The dataset for object detection consist of 2 classes' viz. knife and guns with 300 images and the dataset was split in 80:20 ratios for train and test part.

The Single-Shot Multi-Box Detection (SSD) network is designed to detect objects in a single shot. SSD is a single convolutional network that learns to anticipate and classify bounding box positions in a single pass. The model is trained on a PC with GPU using tensorflow object detection API. Tensorflow offers a diverse choice of models for pre-trained dataset models such as the COCO dataset, the Kitti Dataset, the Open Image Dataset etc. The model has been trained using the following steps as shown in figure 3.

Certain parameters were used to evaluate the performance of the employed model such as loss, training time, mean average precision (MAP).

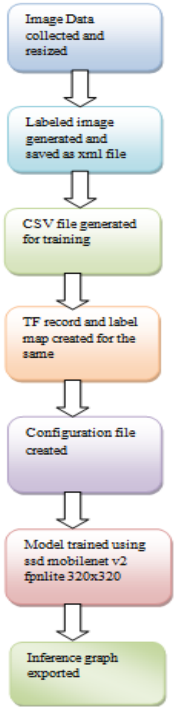


Image Data collected and resized
Labeled image generated and saved as xml file
CSV file generated for training
TF record and label map created for the same
Configuration file created
Model trained using ssd mobilenet v2 fpnlite 320x320
Inference graph exported

Figure 3. Object detection training steps

3.4 Hardware Configuration

Pi camera with raspberry is installed at edge node as shown in figure 4 for streaming live video to server placed at fog end using Wi-Fi-(IEEE 802.11x). The Raspberry Pi NoIR Camera V2 Module consists of 8-megapixel Sony IMX219 image sensor that was specifically designed for Raspberry Pi Kits. The camera lens 'No Infra Red' filter makes it suitable for night and low light photography. Raspberry Pi3 model B (RPi) has a built-in 802.11 b/g/n Wireless LAN and a 400 MHz Video Core IV on its quad-core 64-bit ARM Cortex A53 running Raspbian stretch (clocked at 1.2 GHz).

Buzzer with SIM800l GSM module serves as alert unit subsequent to receiving signal from raspberry pi based on detected unscrupulous activity and weapons. GSM module is a complete quadband GSM/GPRS solution in smt format that can be integrated into customer applications. Quad-band 850/900/1800/1900 MHz is supported by the SIM 800L.

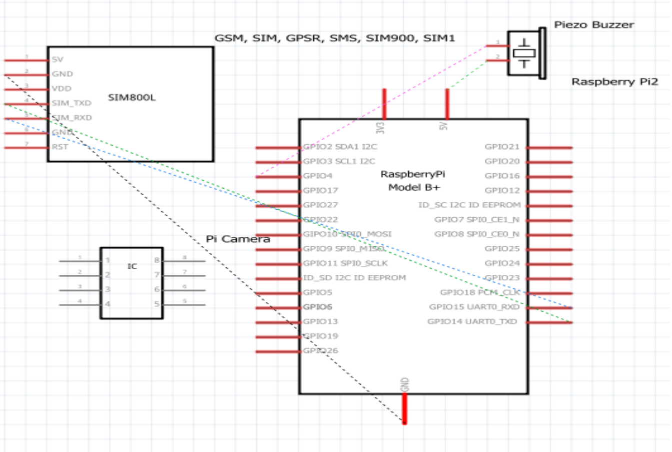


Figure 4. Hardware Connection

3.5 IoT Architecture designed

The study employs the 3 layer IoT architecture viz. physical layer, Fog layer and cloud layer as shown in figure 5. Since the distance between sensors and server in this work is within few meters, medium range communication technology viz. Wifi is the deployed one for the task. Thus WLAN is used as the communication protocol.

i) Physical layer: it consists of things such as sensors that are used to collect information and push it to network layer. This layer is hosting the edge node devices.

ii) Fog layer- this layer hosts the server for processing and returning the output of the supplied input of physical layer via network layer. In comparison to the traditional IoT framework, this layer being closer to the 1st layer is the most distinguishing feature of an IoT service framework in this study with fog computing (FC), as It significantly reduces data processing time by resolving the issues of low bandwidth and severe delivery delay.

iii) Cloud layer: the cloud layer serves the function of storing, updating or can carry out deployment task if needed.

TABLE 1 POSE CLASSIFICATION RESULT

Model	Precision (Mean)%					Recall (Mean)%					F1-Score (Macro average)%	Accuracy (Mean)%
	walking	running	fighting	Laying	Jump	walking	running	fighting	Laying	jump		
Random forest	100	97	95	97	100	99	100	100	100	87	97.54	98.07
Xgboost	100	100	99	100	100	100	100	100	100	94	99.51	99.61
SVM	97	93	96	98	98	97	97	97	99	85	95.66	96.24
DecisionTree	96	95	94	96	96	94	96	95	99	89	95.016	95.18

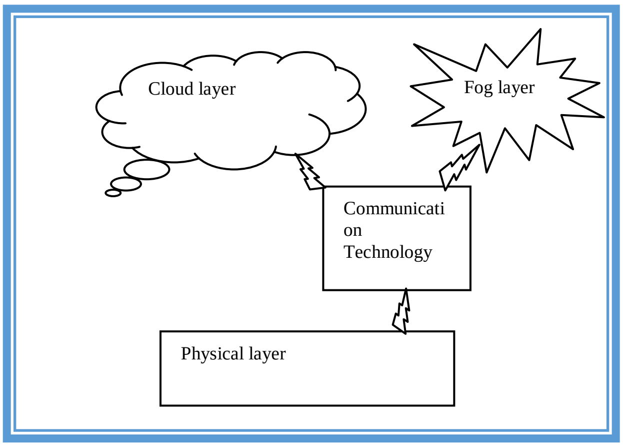


Figure 5. IoT architecture

Video processing is done at edge node and pose estimation, activity recognition and objects detection at fog node. Finally the alerts containing type of abnormal activity, weapons detected if any with distance from the camera are sent using sim800l gsm module placed at edge node via signal sent from fog node to raspberry pi. Further the moment of abnormal activity detection can be snapped and stored at the cloud location for future use.

4 Results and Analysis

Pose estimation model MoveNet with lightning version running at 30FPS works soundly and the extracted-drawn keypoints perfectly resembles the skeleton and synchronizes with the human joints movement as can be seen in figure 2 and figure 7.

4.1 Machine learning models analysis

1) Machine learning (ML) models such as decision tree, random forest, xgboost classifier, and support vector machine were trained using the generated dataset.

2) Though all the models work perfectly after training as can be seen from table 1 but xgboost with ‘gbtree’ booster, ‘mlogloss’ evaluation metrics and ‘multi:softmax’ for multi classification etc outperforms the random forest.

3) Xbgoost classifier achieves 100% score in most of the metrics in terms of precision and recall (table 1). It achieves an accuracy of 99.61% and f1 score of 99.51%.

4) Random forest shows excellent results with accuracy of 98.07% and f1 score 97.54%.

5) Support vector machine (SVM) classifier and Decision tree also shows good results with mean accuracy 96.24 (SVM) and 95.18% (Decision Tree) and macro average F1 score 95.66% (SVM) and 95.016% (Decision Tree).

6) The results are displayed in table 1 and confusion matrixes are plotted in figure 8 for all the models. Figure 10 a) and b) shows accuracy and F1 score graph for all models.

4.2 Object Detection Model

a) A total of 20000 steps were executed for training the model and loss and mean average precision and recall for evaluating the model has been computed as shown in table 2.

b) ssd_mobilenet_v2_fpnlite_320x320_coco17_tpu-8 model was trained for 55 minutes on a gpu graded computer using tensorflow object detection API.

c) Mean average precision (mAP) measures the performance for carrying out object detection tasks. Higher precision means more confidence in model when it classifies a case as positive and higher recall means higher number of cases correctly identifies as positive.

d) The mAP and mAR are tabulated in table 2 and the loss value in table 3. Further figure 6 shows the different loss curves such as classification loss, localisation loss, total loss as well as regularisation loss and the combine pose estimation-classification with object detection is shown in figure 9.

e) Loss portrays the correctness of the model; the minimum loss reaches below 0.25 and was not dipping further, so training was stopped at the mentioned number of steps.

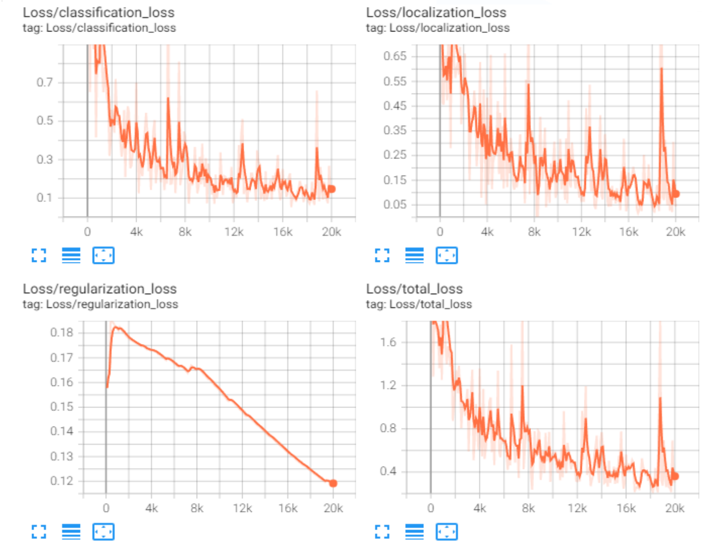


Figure 6. Loss curves

Table 2 mAP and mAR values at different IoU

Model	**mAP IoU = 0.50:0.95**	**mAP IoU = 0.50**	**mAP IoU = 0.75**	**mAR (IoU@0.5:0.95) maxdet=100**
ssd mobilenet v2 fpnlite	0.280	0.620	0.185	0.447

Table 3 Loss parameters

Model	Min loss	Max Loss	Avg loss
ssd_mobilenet_v2_fpnlite	0.129	2.5730	0.8678

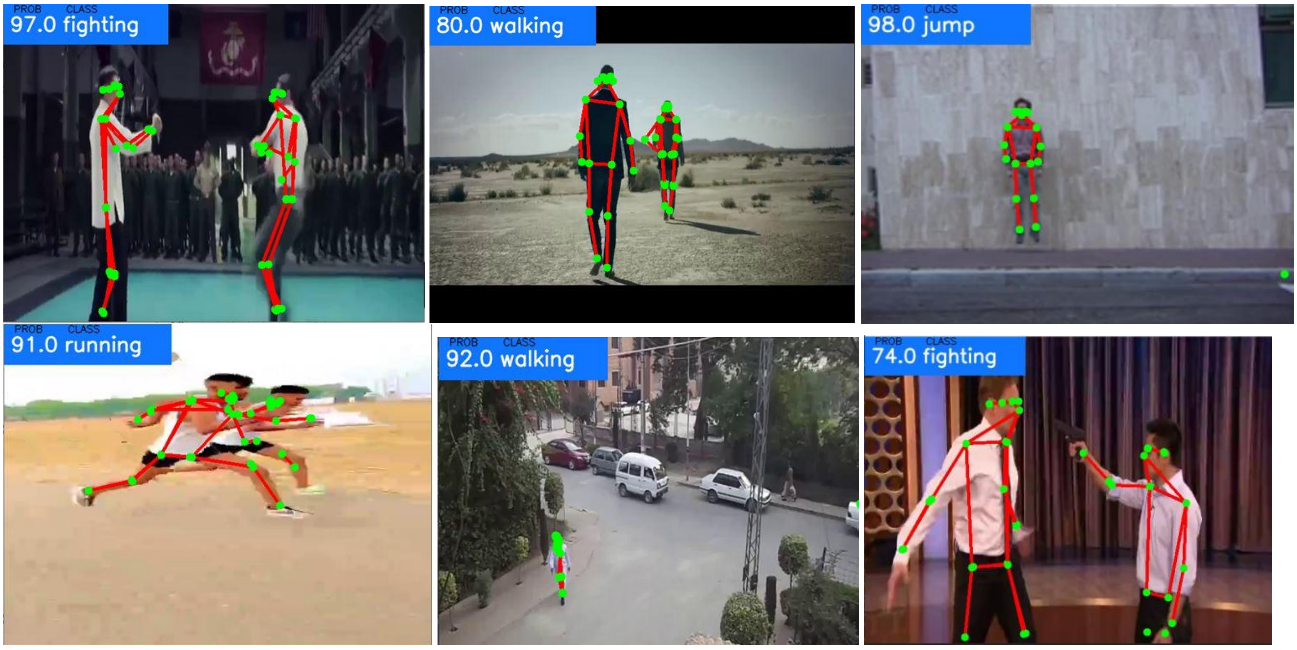
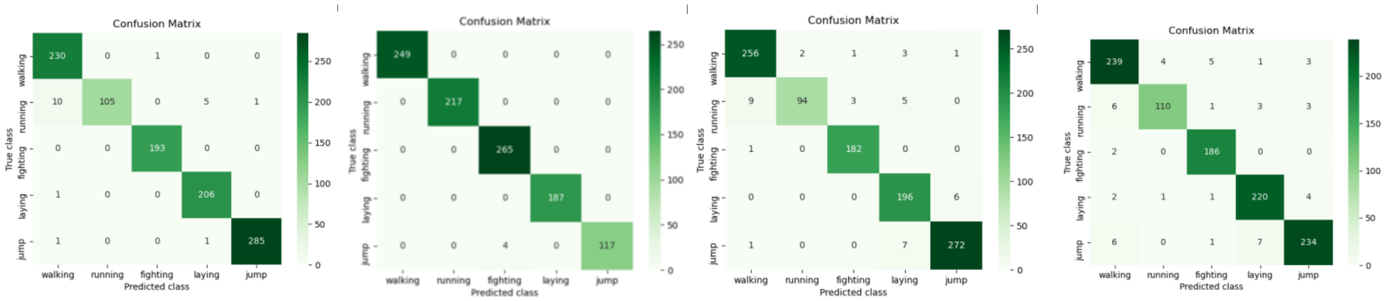


Figure 7. Pose estimation Results



a)Random forest b)Xgboost c)Support Vector Classifier d)DecisionTree

Figure 8. Confusion Matrix’s

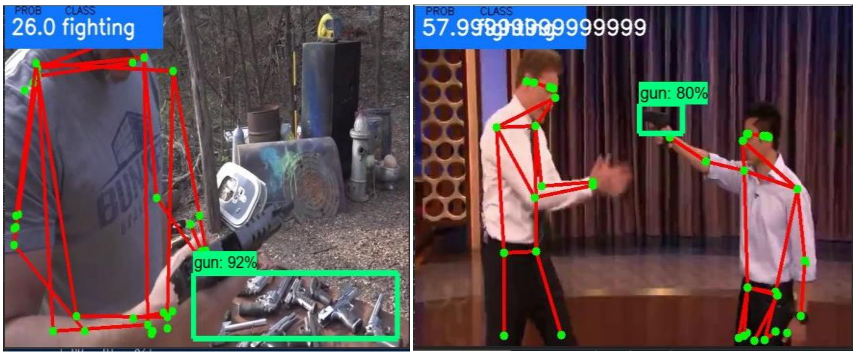


Figure 9. Combine Pose estimation with object detection

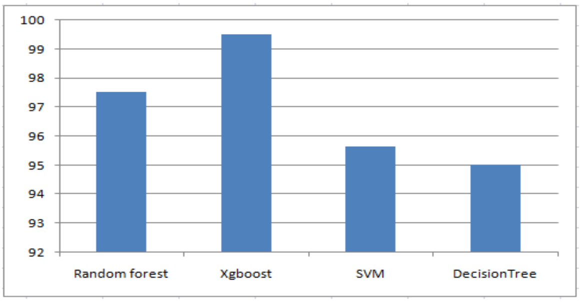


Figure 10(a). Accuracy Graph (Y axis-accuracy, X axis-Models)

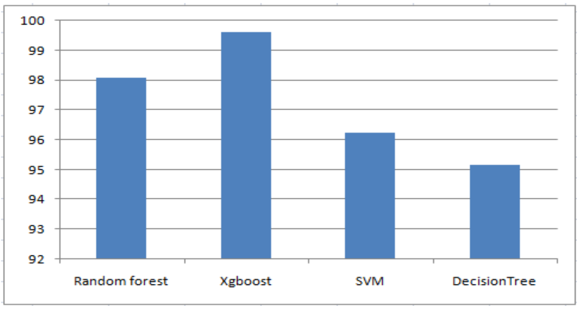


Figure 10b). F1 score Graph(Y axis-F1 score, X axis-Models)

After pose classification and object detection, the server at fog end notifies the raspberry pi to send alerts in the form of sms if unscrupulous activity detected.

5. Challenges

There are few known challenges that need some practical steps. Occlusion can prevent keypoints from detection while subjects at greater distance can get undetected, this can be overcome by installing multiple cameras. But in this work, the major challenge was of utilizing object detection with pose estimation and activity recognition part. Since Object detection based on tensorflow is resource consuming. The FPS (frame per second) which is around 30 while performing pose estimation got reduced to 27-28 while doing pose estimation with activity classification, got further reduced to around 15-16 after including object detection for the employed server. Thus object detection slows the output results though not much effect due to selection of ssd mobile net model as it does the object detection task in one pass but still it may cause some delay not suitable for some specific monitoring. More advanced server can solve this FPS issue but this would increase the cost of operation.

Proposed pose estimation with activity classification using ML models is sufficient for the task of categorizing human activity due to its overwhelming performance. But to further introduce the concept of using object detection task with it and to further hone the result by detecting presence of specific-required object, object detection is used.

Tensorflow lite (tflite) model can be utilized to overcome the problem of speed and all the above task can be made to run on raspberry pi by converting the tensorflow object detection into tflite as it is a framework that helps run models on mobile, embedded, and IoT devices.

6. Discussion

This smart human activity recognition system offers modern solutions in place of manually monitored 24*7 systems. The data was collected from various angles and with varying numbers of subjects, the proposed IoT based armour system works with diverse video inputs. The system presents an IoT enabled novel combination of posture detection, activity recognition with object detection for real time surveillance constituting an IoT based armour protection. With more variety of case dataset such as sitting, doing exercise, the proposed IoT based armour can be made more generic.

Machine learning models (ML) have been thoroughly tested and best one has been deployed. More functionality in the form of adding face detection in object detection can be appended, giving more scope to the system. This IoT based armour can be utilized both at small scale and large scale.

7. Conclusion

Human activity recognition is a challenging research task. This study proposes person and view independent activity recognition using pose estimation and classification techniques. Further object detection has been incorporate to further hone the classification results with the challenges being addressed in using object detection task. The employed ML models show excellent result reaching above 99% of accuracy (XGboost) in classifying activities from diverse camera angles.

References

[1] Antonio Carlos Cob-Parro , Cristina Losada-Gutiérrez , Marta Marrón-Romera , Alfredo Gardel-Vicente and Ignacio Bravo-Muñoz, “smart video surveillance system based on Edge Computing,” *Sensors* 2021, *21*(9), 2958; https://doi.org/10.3390/s21092958

[2] Imran Ullah Khan, Sitara Afzal and Jong Weon Lee, “Human Activity Recognition via Hybrid Deep Learning Based Model” , Sensors 2022, 22(1), 323; https://doi.org/10.3390/s22010323

[3] Daniel O˜noro-Rubio, Roberto J. L´opez-Sastre, Carolina Redondo-Cabrera and Pedro Gil-Jim´enez, “The challenge of simultaneous object detection and pose estimation: a comparative study,” Image and Vision Computing, Elsevier, https://doi.org/10.1016/j.imavis.2018.09.013

[4] Ali Varamesh, Tinne Tuytelaars, “Mixture Dense Regression for Object Detection and Human Pose Estimation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, https://doi.org/10.48550/arXiv.1912.00821

[5] Ming Cheng, Kunjing Cai, Ming Li, " RWF-2000: An Open Large Scale Video Database for Violence Detection, ICPR2020 international conference, https://doi.org/10.48550/arXiv.1911.05913

[6] Josip Josifovski, Matthias Kerzel, Christoph Pregizer, Lukas Posniak, Stefan Wermter, ‘Object Detection and Pose Estimation based on Convolutional Neural Networks Trained with Synthetic Data’, 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Madrid, Spain

[7] Anna Ferrari, Daniela Micucci, Marco Mobilio & Paolo Napoletano, "Deep learning and model personalization in sensor-based human activity recognition," Journal of Reliable

Intelligent Environments https://doi.org/10.1007/s40860-021-00167-w

[8] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 1137–1149.

[9] Malek Al-Nawashi, Obaida M. Al-Hazaimeh & Mohamad Saraee, "A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments," Neural Computing and Appl. 2017, 28, 565-572.

[10] Farzan Majeed Noori, Benedikte Wallace, Md. Zia Uddin and Jim Torresen, "A Robust Human Activity Recognition Approach Using OpenPose, Motion Features, and Deep RecurreNeural Network," Springer SCIA 2019: Image Analysis pp 299-310.

[11] Shuvo Kumar Paul, Muhammed Tawfiq Chowdhury, Mircea Nicolescu, Monica Nicolescu,David Feil-Seifer, "Object Detection and Pose Estimation from RGB and Depth Data for Real-time, Adaptive Robotic Grasping", advances in computer vision and computational biology springer 2021

[12] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, Vitoantonio Bevilacqua," Computer vision and deep learning techniques for pedestrian detection and tracking: A survey", Neurocomputing 2018, 300, 17–33.

[13] Alejandro Newell, Kaiyu Yang, and Jia Deng, "stacked hourglass networks for human pose estimation", european conference on computer vision ECCV 2016: Computer Vision – ECCV 2016 pp 483–499

[14] Velastin, S.A. y Gómez-Lira, D.A. (2017). People Detection and Pose Classification Inside a Moving Train Using Computer Vision. In Advances in Visual Informatics. Lecture Notes in Computer Science, 10645, pp. 319-330.

[15] **https://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html**

[16] M. B. Holte, C. Tran, M. M. Trivedi and T. B. Moeslund, "Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments," in IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 5, pp. 538-552, Sept. 2012, doi: 10.1109/JSTSP.2012.2196975.

[17] P. Yang, C. Yang, V. Lanfranchi and F. Ciravegna, "Activity Graph Based Convolutional Neural Network for Human Activity Recognition Using Acceleration and Gyroscope Data," in IEEE Transactions on Industrial Informatics, vol. 18, no. 10, pp. 6619-6630, Oct. 2022, doi: 10.1109/TII.2022.3142315.

[18] Ali S, Shah M. Human action recognition in videos using kinematic features and multiple instance learning. IEEE Trans Pattern Anal Mach Intell. 2010 Feb;32(2):288-303. doi: 10.1109/TPAMI.2008.284. PMID: 20075459.