**Intelligent Intrusion Detection Using TensorFlow**

Adarsh Padhi2 , Gyanaranjan Sahoo2 , Satyabrata Maity1

Computer Science & Engineering12, Siksha ‘O’ Anusandhan University, Odisha, India12

[satyabratamaity@soa.ac.in](mailto:satyabratamaity@soa.ac.in)1, {adarshpadhi14, gyanaranjan7205}@gmail.com2

**Abstract:** Detecting human beings accurately in a visual surveillance system is crucial for diverse application areas including abnormal event detection, human gait characterization, congestion analysis, person identification, gender classification and fall detection for elderly people. The first step of the detection process is to detect an object which is in motion. Object Detection [1] is the task of identifying the presence of predefined types of objects in an image. This task involves both identification of the presence of the objects and identification of the rectangular boundary surrounding each object (i.e. Object Localization). An object detection system which can detect the class “Human” can work as a Human Detection System. To overcome the difficulties of the current scenarios where people are wasting much of time and resources in detecting intrusions; we proposed a solution utilizing Machine Learning [2] with help of TensorFlow [3] which could get us out of this hefty situation by trimming videos containing suspicious activity and also act intelligently during surveillance.

The processed video could possibly shorten the time and memory data consumption. It could also help in searching suspicious or malicious activities easily.

**Keywords:** TensorFlow, Machine Learning, Object Detection, Video Processing.

1. **Introduction**

Most of us in our daily life, are worried about the thieves, accidents, murders, etc. Everyone is aware of this and have love for their life. So, they appoint Watchmen, Security Guards, Bodyguards in order to protect their lives’. Now a day’s people were much serious with their belonging. So, every were a CCTV camera is being installed in order to capture all the things that is a beyond human capability.

Incase,

1. There is a robbery and the user don’t have sufficient time and patience to review it.
2. There is less storage space available for storing CCTV videos.
3. There is fire in an unattended house or is about to have any lethal accident due to weapons.

We have overcome above the situation, by Object Detection it could really helpful and can bring a change in our society. We have designed a software that can possibly overcome the problems mentioned above.

Machine Learning is a field of computer software study in which the user train the computer or simply we can say that It makes the computer to think as well as learn. It is a part of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Machine learning algorithms are often categorized as supervised or unsupervised. Object recognition is a general term to describe a collection of related computer vision tasks that involve identifying objects in digital photographs. Image classification involves predicting the class of one object in an image. Object localization refers to identifying the location of one or more objects in an image and drawing abounding box around their extent. Object detection

***Motion Detection*** is the process of detecting a change in the frame of an object relative to its surroundings or a change in the surroundings relative to an object. It also implements ***Face Detection*** which is a computer technology being used in a variety of applications that identifies human faces in digital images. It also refers to the psychological process by which humans locate and attend to faces in a visual scene.

In this paper, we will be explaining the basics and applicability of Human Detection and the effectiveness of this technology for taking valid frame footage, sharing of data through messages and online streaming which were required for the user to be alert and can take precaution before dealing with the situation.

1. **Related Work**

Manoranjan Paul [4] described a number of ways utilized in Object Detection such as background subtraction which detects object by difference in current and reference frame in block by block fashion, optical flow approach which estimates motion in video by matching points over image frames, spatio-temporal filter are better to capture both spatial and temporal information of gait motion.

M. M. Sardeshmukh [5] used background subtraction methods to detect human and track in different levels of illumination. The foreground is captured by subtracting previous frames from a test frame by checking pixel values.

Muhammad Attique Khan [6] introduces a method for human detection and action recognition which combines Euclidean distance and joint entropy-based features selection and uniform segmentation. It segments the moving objects in a frame by fusing novel uniform segmentation and expectation maximization, then its extracts a new set of fused features using local binary patterns with HOG and Haar-like features. The features are selected by novel Euclidean distance and joint entropy-PCA-based method and is classified using a multi-class support vector machine.

Javier Oliver [7] utilize HAAR-Boosting-based methods for moving human detection. He named the method as Haar-Boosting-HOG-SVM detection method to detect pedestrians. In this method Haar-like features are extracted from a frame and is classified using Boosting based method.

1. **Proposed Solution**

We have designed a software which will reduce the time as well as the data consumption rate. The project “Intelligent Intrusion Detection” perfectly defines itself. In this project we have used Machine learning which enables to detect between humans and all other species and objects. We have used TensorFlow model with MSCOCO [8].

TensorFlow is a [free](https://en.wikipedia.org/wiki/Free_software) and [open-source](https://en.wikipedia.org/wiki/Open-source_software) [software library](https://en.wikipedia.org/wiki/Library_(computing)) for [dataflow](https://en.wikipedia.org/wiki/Dataflow_programming) and [differentiable](https://en.wikipedia.org/wiki/Differentiable_programming) programming across a range of tasks. It is a symbolic math library, and is also used for [machine learning](https://en.wikipedia.org/wiki/Machine_learning) applications such as [neural networks](https://en.wikipedia.org/wiki/Neural_networks) and , MSCOCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features: Object segmentation. Recognition in context.

The detection concept is basically based on the Neural Networks basically RCNN [9] i.e., (Region Based Convolution Neural Networks). It generates bounding boxes and segmentation masks for each instance of an object in the image. It's based on Feature Pyramid Network [10] (FPN) and a ResNet101 backbone.

The application takes a video or webcam stream as an input and writes out the frames containing any human intrusion, thereby making a brief video. This results to shorten of original video and data consumption rate which leads to save of time and storage space.

This project also plays a vital role in protection, due to its double layered security. Firstly, when it finds the intrusion, the first level of security gets activated and it sent a Message or a call to the owner about informing that there is some suspicious in your house. If the application detects weapons like (knife, guns, fire) then the 2nd Level of security gets activated and it sends a message to owner and on confirmation or no response from owner for over 10 minutes it sends location and an alert message to the nearby Police Station.

The application also supports Live Streaming of webcam or video input over Internet to platforms like YouTube and Twitch.

1. **Prototype**

Human detection in a smart surveillance system aims at making distinctions among moving objects in a video sequence. The successful interpretations of higher level human motions greatly rely on the precision of human detection.The detection process occurs in two steps: object detection and object classification by deep learnng techinques. Among the deep learning based techniques, two broad class of methods are prevalent: two stage detection (RCNN, Fast RCNN, Faster RCNN ) and unified detection (Yolo, SSD).

The network used in this project is based on Single shot detection (SSD) [7]. The architecture is shown below:



The SSD normally starts with a VGG model, which is converted to a fully convolutional network. Then we attach some extra convolutional layers, that help to handle bigger objects. The output at the VGG network is a 38x38 feature map (conv4 3). The added layers produce 19x19, 10x10, 5x5, 3x3, 1x1 feature maps. All these feature maps are used for predicting bounding boxes at various scales (later layers responsible for larger objects). Thus the overall idea of SSD is shown below. Some of the activations are passed to the sub-network that acts as a classifier and a localizer.

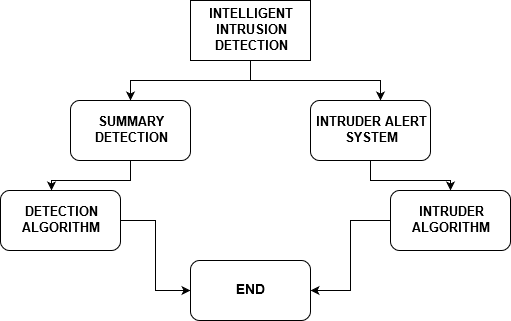


Anchors (collection of boxes overlaid on image at different spatial locations, scales and aspect ratios) act as reference points on ground truth images. A model is trained to make two predictions for each anchor:

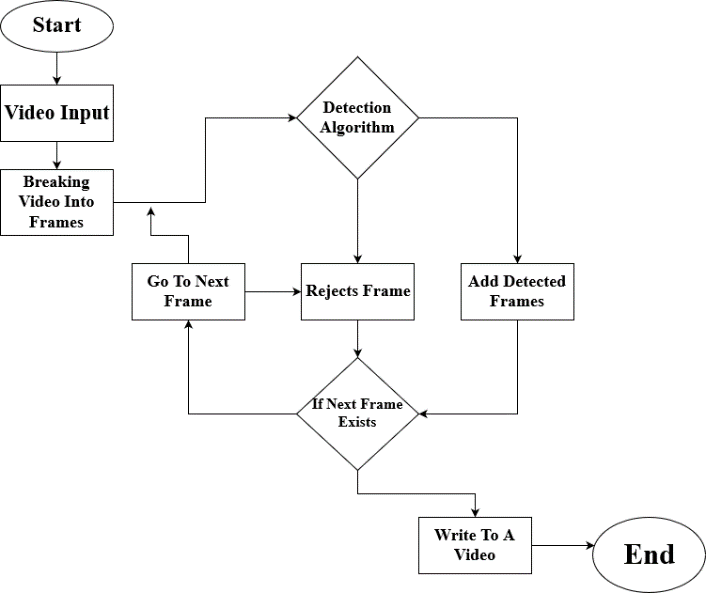
* A discrete class
* A continuous offset by which the anchor needs to be shifted to fit the ground-truth bounding box.

During training SSD matches ground truth annotations with anchors. Each element of the feature map (cell) has a number of anchors associated with it. Any anchor with an IoU (jaccard distance) greater than 0.5 is considered a match. The loss function used is the multi-box classification and regression loss. The classification loss used is the softmax cross entropy and, for regression the smooth L1 loss is used. The SSD model consists of the base network derived from RESNet and then the modified convolutional layers for fine-tuning and then the classifier and localizer networks. This creates a deep network which is trained end-to-end on the dataset.

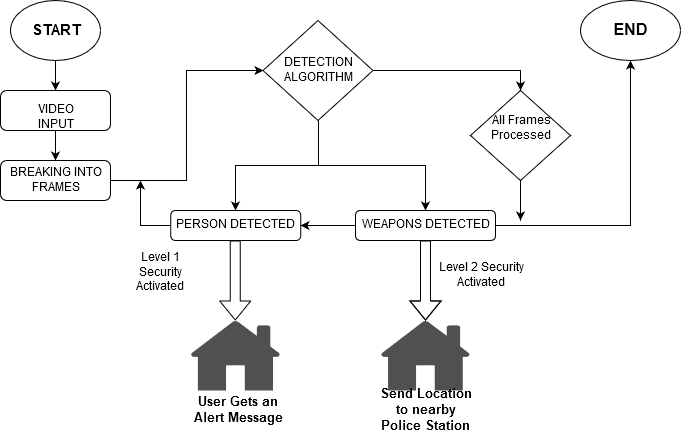
The project can be used in 2 different modes exclusive to each other i.e. storage mode, surveillance mode.



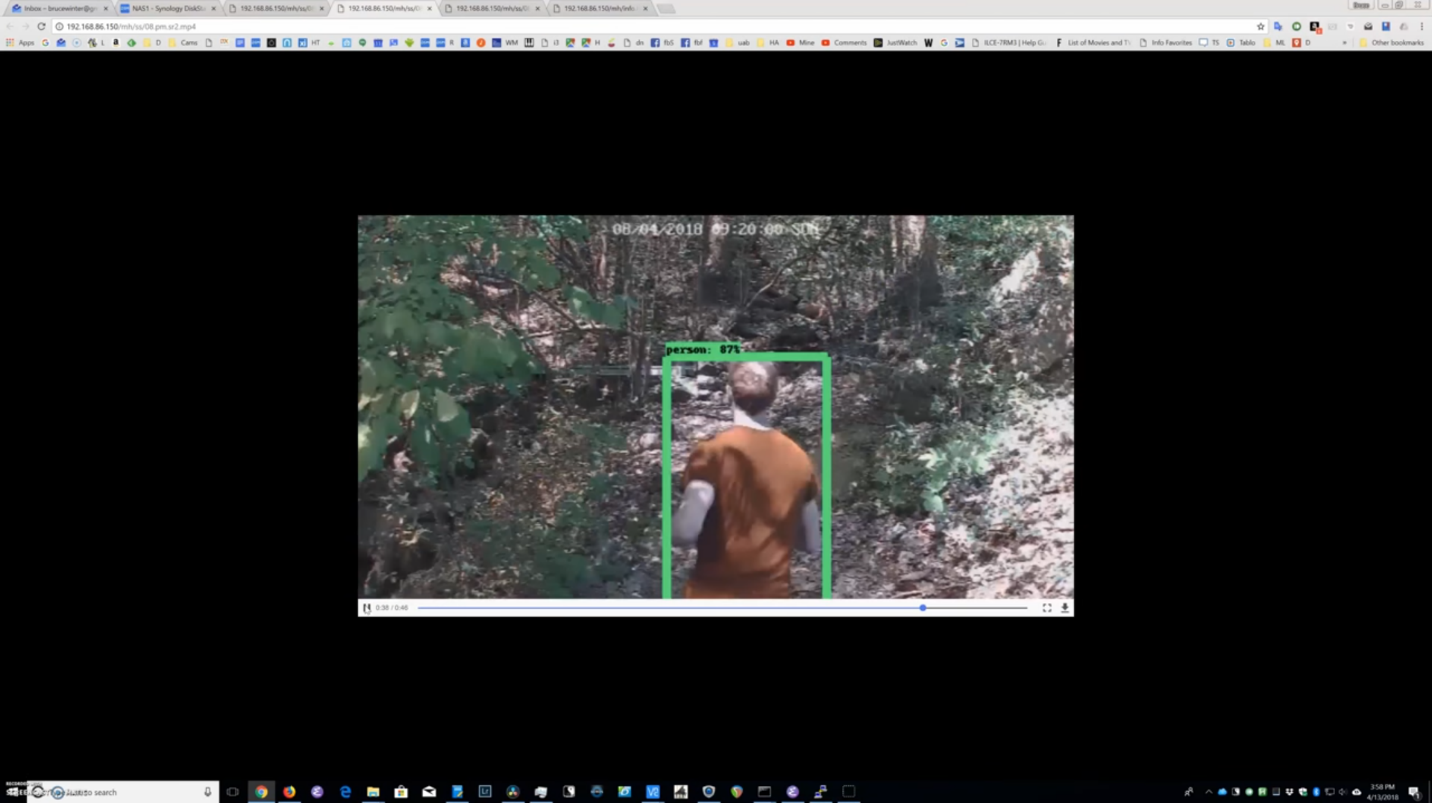
Storage Mode: This mode writes only those frames in which any human intrusion is detected to a video file along with current timestamp. The video source can either be a video file or a live webcam. This will help user in several ways like – producing a brief output video such that it can be reviewed fast thereby saving time, taking a much smaller storage space than the original video should have taken thereby saving resources and storage space.



Surveillance Mode: Under this mode, the application actively takes part in security of user. It has 2 levels of security. In 1st level, it informs the user when it finds any human locomotion in its area of vision by either a SMS or a call. On 2nd level, when found any weapons like knife, daggers, guns or fire it informs the user by a message with a link to deactivate the alarm if the alarm is not deactivated in 10 minutes the location along with an alert message is sent to nearest Police Station. SMS / Call are sent by utilizing API provided by Gateway. Live Stream of video to websites like YouTube and Twitch was also made possible by using ffmpeg library.



Glances of detection from application:



Test on Data Samples:

A short video of a burglary on road was taken from YouTube. From the video 150 frames were taken out in total at regular interval. Four such tests were done and results were manually verified. The results obtained were classified into 4 parts which are

Total Positive: When the experimental results show that a frame **contains** a human and in actual data there **is** a human, it is considered to be a Total Positive case.

False Positive: When the experimental results show that a frame **contains** a human and in actual data there **is no** human, it is considered to be a False Positive case.

False Negative: When the experimental results show that a frame **does not contains** a human and in actual data there **is** a human, it is considered to be a False Positive case.

Total Negative: When the experimental results show that a frame **does not contains** a human and in actual data there **is no** human, it is considered to be a False Positive case.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Count | True Positive | False Positive | False Negative | True Negative |
| Video Segment I | 82 | 3 | 26 | 39 |
| Video Segment II | 77 | 1 | 24 | 48 |
| Video Segment III | 104 | 0 | 34 | 12 |
| Video Segment IV | 63 | 1 | 25 | 61 |

|  |  |  |  |
| --- | --- | --- | --- |
| Test Count | Specificity  TN/(TN+FP) | Precision  TP/(TP+FP) | Recall  TP/(TP+FN) |
| Video Segment I | 0.9286 | 0.9647 | 0.7593 |
| Video Segment II | 0.9796 | 0.9872 | 0.7624 |
| Video Segment III | 1 | 1 | 0.7536 |
| Video Segment IV | 0.9839 | 0.9844 | 0.7159 |

The project can be used for several purposes and has several use cases. Some of them are:

It can be used in active or passive mode for saving storage space by saving only those frames when human is detected and live stream them.

It can be used in research purposes (in places where u wait for an event to happen) to inform you when that event happening like monitoring tigers near forest areas, etc.

The application can work intelligently to detect any causalities like fire, accidents, robbery and put on alarm or inform the owner also the police station so that necessary actions can be taken.

**6. Conclusion**

A significant amount of work has been done with a view to detect human beings in a surveillance video. However, the low-resolution images from the surveillance cameras always make this work challenging. From the machine vision perspective, it is hard to distinguish an object as a human due to its large number of possible appearances. Moreover, the motion of human is non periodic in nature. But a combination of several features could be useful identifying human. The application can be further improved to work with multiple cameras and process frames of each time frame from different angles to provide better results, accuracy and lesser false negatives. Although the application is not 100% accurate still it has a better accuracy in comparison with speed and memory usage. This application could be a great help to people along with push the current real-world implementation techniques a bit nearer to the World of Automation.

The project can be further improved in various ways and will be considered as future work. Some of them are – First, detecting abnormal events like single-person loitering, multiple-person interactions (e.g. fighting and personal attacks), person-vehicle interactions (e.g. vehicle vandalism), and person-facility/location interactions (e.g. object left behind). Second, person detection in dense crowds and people counting and pedestrian detection which could be helpful in a variety of other application. Third, tracking a specific person in a visual surveillance system by utilizing using face recognition and gait recognition techniques. Fourth, automatic detection of a fall for elderly people is one of the major applications which can be implemented to human detection in surveillance videos.

1. **References**

[1] Amit Y., Felzenszwalb P. (2014) Object Detection. In: Ikeuchi K. (eds) Computer Vision. Springer, Boston, MA

[2] Liu, L., Ouyang, W., Wang, X. *et al.* Deep Learning for Generic Object Detection: A Survey. *Int J Comput Vis* (2019) doi:10.1007/s11263-019-01247-4

[3] Toleubay Y., James A.P. (2020) Getting Started with TensorFlow Deep Learning. In: James A. (eds) Deep Learning Classifiers with Memristive Networks. Modeling and Optimization in Science and Technologies, vol 14. Springer, Cham

[4] Paul, M., Haque, S.M.E. & Chakraborty, S. EURASIP J. Adv. Signal Process. (2013) 2013: 176. <https://doi.org/10.1186/1687-6180-2013-176>

[5] Sardeshmukh M.M., Kolte M., Joshi V. (2016) Performance Analysis of Human Detection and Tracking System in Changing Illumination. In: Corchado Rodriguez J., Mitra S., Thampi S., El-Alfy ES. (eds) Intelligent Systems Technologies and Applications 2016. ISTA 2016. Advances in Intelligent Systems and Computing, vol 530. Springer, Cham

[6] Sharif, M., Khan, M.A., Akram, T. et al. J Image Video Proc.(2017) 2017:89. https://doi.org/10.1186/s13640-017-0236-8

[7] Oliver J., Albiol A., Morillas S., Peris-Fajarnés G. (2009) A Real-Time Person Detection Method for Moving Cameras. In: Corchado E., Wu X., Oja E., Herrero Á., Baruque B. (eds) Hybrid Artificial Intelligence Systems. HAIS 2009. Lecture Notes in Computer Science, vol 5572. Springer, Berlin, Heidelberg

[8] Lin TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham

[9] Yamashita, R., Nishio, M., Do, R.K.G. et al. Insights Imaging (2018) 9: 611. https://doi.org/10.1007/s13244-018-0639-9

[10] He K., Zhang X., Ren S., Sun J. (2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8691. Springer, Cham