Artificial Intelligence

14

AI in Video Analytics and Security

Punna Rajasekhar*

Security Electronics & Software Systems Division, Bhabha Atomic Research Centre, Mumbai 400085, INDIA

Zones are marked up to 50m. Red line is virtual fence and green is zone

ABSTRACT

The surveillance of critical infrastructures and installations are typically based on multiple video cameras via a centralised manual monitoring station, among other modalities. The continuous monitoring of the video inputs from multiple cameras by human is error prone due to fatigue. Assistance from automated systems for detecting various malicious activities and movements can increase the surveillance performance. Due to the huge success of Artificial Intelligence (AI) based systems in analysing the images for various computer vision tasks, the same can be extended for analysing live surveillance videos. This article provides details of the AI based video analytics systems under development at SESSD, BARC.

KEYWORDS: *Deep learning, Computer vision, Video analytics, Object detection, Stereo vision*

Introduction

Video cameras are found inevitably in all modern surveillance systems of all critical installations and infrastructures. However, monitoring the feed from all these large number of video cameras at a centralised monitoring/operator room with a video wall is tedious and error-prone task. Studies[1] have shown that a human operator, watching a single video monitor for more than 20 minutes loses 95% of his ability to pay sufficient attention to differentiate between normal and malicious events. This has, hence, largely restricted the usage of this highly rich live data content only to be used during post-event analysis of the recorded video feed, in most scenarios. This has inspired the need to have automated solutions for analysing the live surveillance video feeds for detection of malicious events viz., loitering, crowd gathering, entering exclusion zones, abandoning baggage etc.

The application of artificial intelligence (AI) algorithms, especially Deep Neural Networks (DNNs), in the fields of image processing/computer vision has drastically improved after 2012. DNNs have been extensively studied and applied for various computer vision tasks like image classification, object detection, pose estimation etc. and has even achieved human level accuracies in few of these tasks. The availability of the high computation hardware like GPUs and the availability of large datasets for computer vision tasks like ILSVRC[2] have enhanced the development of deep learning techniques in recent years. Relying on the huge success of deep learning in analysing the images, it is possible to extend the same method to analyse video stream for detecting malicious events in real time. This can help in assisting the security personnel in surveillance and reduce the errors caused due to fatigue.

SESSD, BARC is involved in the development of various deep learning based automated system for assisting surveillance. These systems will be deployed for field testing within BARC. The Deep Learning based systems under development are as below:

1) Physical Intrusion Detection System
   a) Stereo Camera based
   b) Single Camera based
   c) Virtual Fence
2) Suspicious Behaviour Detection System

Each of these systems are briefly described in the following sections.

Deep Learning based Video Analytics Surveillance Systems

Many surveillance cameras are already installed along the fence/boundary of a critical installation. Typically, the modern-day cameras are digital IP based cameras and hence the video feed from all these cameras is available at a centralised monitoring station. The primary objective of the proposed video analytic systems is to identify activities like, any human approaching the fence, crossing the fence, Loitering in an exclusion zone etc.

Physical Intrusion Detection System (PIDS)

PIDS is a vision-based system, which detects human intrusions and also finds the 3D coordinates of the location of the intrusion w.r.t to camera coordinates system (CCS). CCS is defined using Geographic Information System (GIS) with x-axis pointed towards the geographic east, y-axis towards

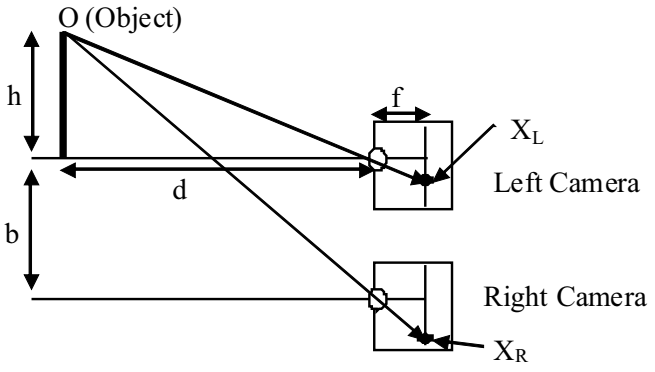

Fig.1: Representation of Stereo camera setup.

*Author for Correspondence: Punna Rajasekhar
E-mail: rajs@barc.gov.in

Fig.2: Image showing a blue coloured bounding box marked over a detected intrusion. The depth, width and height (in meters) of the intrusion are also indicated over the bounding box as text.

geographic north and z-axis pointing up from the ground; camera feet location as origin of CCS. This kind of orientation assumption helps in integrating the live intrusion information from multiple camera feeds on to a single fused map.

In PIDS, the video feed received from each camera is analysed to find the intrusions per-image basis. The objective of finding the intrusions in a frame is realised by a deep learning model trained to detect objects in the given frame. The deep learning model is trained on large publicly available MS-COCO[3] dataset and the model architecture is based on SSD[4]. All the deep learning models are trained using Pytorch framework . Once the 2D-location of the humans is obtained in pixels, it is mapped to 3D coordinate system (CCS). Two different approaches are adopted for 2D to 3D recovery.

Stereo based Approach

In the stereo approach[6], two cameras (denoted as left and right cameras) are placed in a stereo fashion, with parallel optical axes, as shown in Fig.1. Assuming a pin-hole camera model, the depth of any object (O) in 3D space, whose 2D pixel locations are known in both the cameras, (X_L and X_R) can be estimated using the equation-1.

$$d = \frac{(f * b)}{(X_R - X_L)} \quad (1)$$

Where *f* is the focal length of the cameras & *b* is the baseline of the setup. The term $X_R - X_L$ is called disparity i.e., the shift in the pixel location of the corresponding 3D point in 2 images.

This system is deployed for field testing within BARC. Results are shown in the Fig.2 below. The system also has panning capability and hence system rotates to cover a larger field of view. The execution time for processing one frame is around 20msec on a PC with GeForce RTX-2080 graphic card (4352 CUDA cores and 11GB RAM).

Single Camera Approach

A stereo vision based setup requires two cameras which are arranged in precise parallel fashion to estimate the depth. The precise alignment of the cameras is mechanically tedious which increases deployment time. In order to avoid the usage and task of aligning of two cameras, a single camera approach is also developed[7]. Under a valid assumption that human stands on the ground, the 3D coordinates of the human are estimated. Feed from any existing surveillance camera can be utilised to detect intrusions using the trained deep learning model for object detection. The 3D location of the feet of a human is estimated from 2D pixel location of feet.

This system is deployed for field testing within BARC. Results are shown in the Fig.3. It can be seen that the z-coordinate of the human feet location is always zero; this assumption is made because human stands on the ground as shown in Fig.3. The execution time for processing one frame is around 20msec on a PC with GeForce RTX-2080 graphic card (4352 CUDA cores and 11GB RAM).

Virtual Fence

The idea of recovering the 3D coordinates of the points of a plane from their 2D pixel locations, is applied to draw virtual fences. This is shown in Fig.4. This application detects a person and also tracks the person. Here, the zones are virtually

Fig.3: Images showing intrusion with their 3D coordinates of feet and height(in meters) of the intrusion.

Fig.4: Zones are marked up to 50m. Red line is virtual fence and green is zone. It detects persons and tracks them. As soon as the person enter the zone, a bounding box is drawn around him and the tracking of the movement is performed.

marked on the ground, as zone-1 and zone-2. If the person enters into one of the exclusion zones, an alarm is generated. This also detects the zone crossing, if the person moves from one zone to other zone. This is deployed within BARC for field testing.

Suspicious Behaviour Identification System (SBIS)

The objective of this application is to find a suspicious action of a person or other objects from the video feed. An end-to-end deep learning architecture is under development to predict suspicious actions. The spatio-temporal features of the key joints of the human in small video clips are used to identify the actions and to further classify into normal or suspicious action. Other possible way is to model a self-learning AI model to learn the normal and suspicious behaviour of a person.

Other Applications

The model obtained from training on MS-COCO[3] dataset for object detection is fine-tuned to perform object detection on the thermal images. These thermal images of human-beings are obtained from FLIR[8]. This process is called transfer learning. The results are shown below in Fig.5. In areas where there is no/low light, it is possible to detect the objects using thermal images.

Summary

Manually monitoring the video feed from several surveillance cameras for detecting any intrusion or suspicious behaviour is a tedious task for security personnel. Deep Learning based models are yielding promising results in various computer vision tasks, viz, object detection and image classification. This article summarises few surveillance systems which are under development where deep learning concepts are being applied. Object detection models are used in Physical Intrusions Detection System and virtual fence systems. Object tracking models are used in virtual fence. These models are trained using publicly available datasets and then these models are fine-tuned using local datasets for surveillance applications. Few of these systems are deployed at BARC perimeter for field testing.

References

[1] M. Green, "The appropriate and effective use of security in schools," US Department of Justice, 1999.

[2] http://www.image-net.org/challenges/LSVRC/ (accessed Sep 5, 2022).

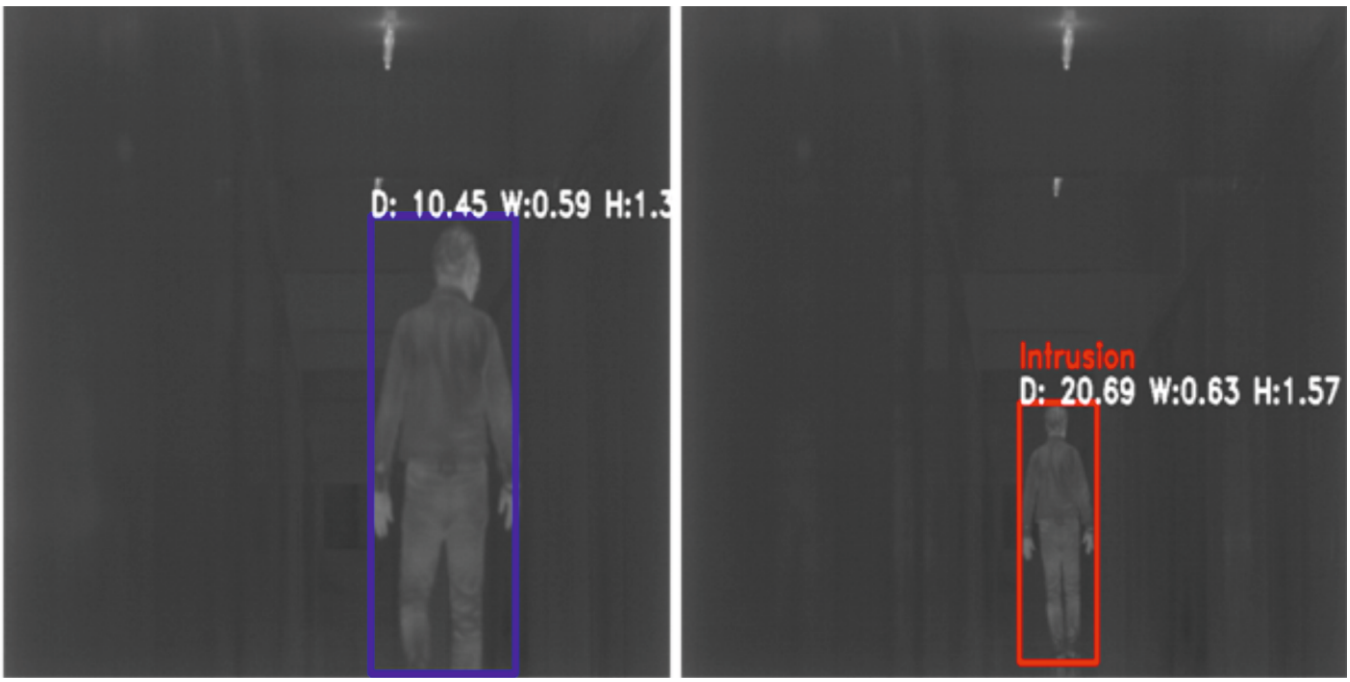

Fig.5: Results of Transfer learning on Thermal Images. The depth, width and height (in meters) of the intrusion are also indicated over the bounding box as text. The experiment was conducted to detect the humans from thermal images placed in stereo-fashion and find the dimensions of the human. Depending on the depth (or distance) of the human from camera, the colour of the bounding box changes. The colour of the bounding box is kept as blue for depths less than 20m and red for depths greater than 20m.

[3] L. T. e. al., "Microsoft COCO: Common Objects in Context," in Computer Vision – ECCV, 2014.

[4] Chengcheng Ning, Huajun Zhou, Yan Song and Jinhui Tang "Inception Single Shot MultiBox Detector for object detection," in IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, 2017.

[5] http://www.pytorch.org (accessed Sep 5, 2022).

[6] Aravamuthan, G., Rajasekhar, P., Verma, R. K., Shrikhande, S. V., Kar, S., Balur, S. "Physical Intrusion Detection System using Stereo Video Analytics," in CVIP-2018-IIITDM Jabalpur, 2018.

[7] R. Punna, G. Aravamuthan and S. Kar, "Recovering 3D from 2D image points (calibrated camera approach for surveillance)," in 2022 IEEE 7th International Conference on Converging Technologies (I2CT), Pune, 2022 .

[8] https://www.flir.in/oem/adas/adas-dataset-form/(accessed Nov 5, 2021).