*Journal of Network Intelligence*
**Taiwan Ubiquitous Information**
 ISSN 2414-8105 (Online)
Volume **6**, Number **3**, August **2021**

# An Intelligent Video Surveillance System Based on Object Recognition and Abnormal Behaviour Detection

Jia-Lin Cui[1], Zhen Wang[2], Long-Hua Ma[1], Zhe-Ming Lu[*3] and Hao-Lai Li[4]

[1]School of Information Science and Engineering
NingboTech University
Ningbo 315100, P. R. China
cuijl_jx@163.com

[2]College of Control Science and Engineering
Zhejiang University
Hangzhou, 310027, China
zwang.lab@zju.edu.cn

[3]School of Aeronautics and Astronautics
Zhejiang University
Hangzhou, 310013, China
*Corresponding Author: zheminglu@zju.edu.cn

[4]EFORT Intelligent Equipment Co., Ltd.
Shanghai 201600, P. R. China

Received March 2021 Revised June 2021;

---

Abstract. *Harbours operating under hazardous conditions, require video surveillance systems for timely discovery of violations and emergencies to avoid serious injuries and even loss of lives. However, traditional monitoring systems guarded manually are manpower consuming and absent easily. To address this difficult problem, we propose an intelligent multi-camera video surveillance system based on micro object detection and abnormal behavior judgments. Our system is able to estimate the location of object and classify them in each frame, then predict whether there is a fire and count the number of people and cars respectively, moreover, predict possible unsafe interactions between people and vehicles. It emphasizes the flexible selection of different algorithm modules in various environments and scenarios. Through the real-time evaluation of the degree of the abnormality of various scenes by our system, the high-scoring situations generate relevant alarms and are preferentially pushed to the dutykeeper. The experiments show that the proposed method detects and evaluates anomalies accurately, schedules monitoring screens more rationally and reduces the workload of staff.*
**Keywords:** Intelligent Multi-Camera Video Surveillance, Micro Object Detection, Real-time, Abnormal behavior.

1. **Introduction.** Video surveillance systems have long been used in a variety of industrial scene. Camera manufacturers are also moving toward intelligent direction. Motion detection and license plate recognition have been embedded in the chip of camera as normal module. However, object recognition and abnormal behavior detection for more complex environments are still at the stage of trial. With the improvement of GPU performance and optimization of parallel computing architecture, detection algorithms based on deep convolutional neural networks are turning to industrial application gradually. The port has a complicated working environment and strict rules and regulations, but

there are still accidents every year, which are generally caused by the illegal operation of workers and the untimely handling of emergency situations. Although the port has already been equipped with a video surveillance system, the degree of intelligence is limited to the identification of the container number, and the ability to alert the abnormal situation is not achieved. For a large industrial environment such as a port, the area is usually hundreds of thousands of square meters, and there are 200 spherical panoramic cameras in the working area. At present, it can only be monitored by human to determine whether abnormal conditions occur. The operator monitors all the camera scenes, picks out dangerous scenes for critical surveillance, and records the violations of workers. Such boring work imposes a great burden on the workers, it is easy to mental fatigue, and there is no guarantee that no abnormal pictures will be missed, and real-time alarms for important alarms cannot be guaranteed. In order to solve the above problems, we propose an intelligent video surveillance system based on convolutional neural networks.

Video surveillance systems have been applied in various areas such as traffic management, crowd behavior detection, and wildlife protection [1]. Traditional object detection algorithms are features-based methods [2], such as texture features. A conmparative research among Discrete Wavelet Transform(DWT), Histogram of Oriented Gradient(HOG) and Speeded Up Robust Feature(SURF) shows that HOG feature with Naive Bayes is better than others [3]. Konda in [4] propose a motion features instead of pixel-based approaches achieving at the same time detection and segmentation. It is effective for moving objects. Manually selected features perform well in applications for human detection and tracking [5]. Recently, neural network is used for human activity recognition in video surveillance [6]. And the multi-target tracking identification system under multi-camera surveillance system [7] has emerged.

A few years ago, the video surveillance system usually used texture features in detection. For complex environments such as port, there are many types of targets to be identified. It is unrealistic to select features for each type of target. Therefore, we use deep convolutional neural network to learn and select features. As long as we prepare a large number of labeled samples, we can achieve object recognition and positioning after training. This paper firstly detects the input video stream frame by frame to obtain the object position. Here, the tracking algorithm is discarded and a pure detection algorithm is used. Because the model is fast enough, real-time detection can be performed on the gpu, and fire alarms and population statistics can be done at the same time. For scenes of people and cars, the system predicts the possibility of an accident and give an alarm.

The rest of the paper is organized as follows. Section 2 is discusses the work related to object detection and abnormal behavior detection.Section 3 explains the algorithm embedded in our system. In Section 4, we describe the experiment to demonstrate the effectiveness of our system. The last section concludes our research and future work.

## 2. **Related Work.**

Infrastructure and algorithm are critical for a complete system. We program the module of taking video steam and visualization by C++, and build a deep learning network model by Python. This can integrates the advantages of C++ running fast and Python's convenient of building a deep learning framework. Then calling the Python code from C++ to compose a complete system.

For an intelligent system, one of the basic problems in video surveillance is object detection, which aims at locating and classifying existing objects in an image. The object detection frameworks can mainly be categorized into two types [8]. One is the region proposal based methods, such as R-CNN [9], Fast R-CNN [10], Faster R-CNN [11], FPN [12] and Mask R-CNN [13]. The other is the regression or classification based methods mainly include YOLO [14], SSD [15], YOLOv2 [16], DSSD [17], DSOD [18], YOLOv3 [19].

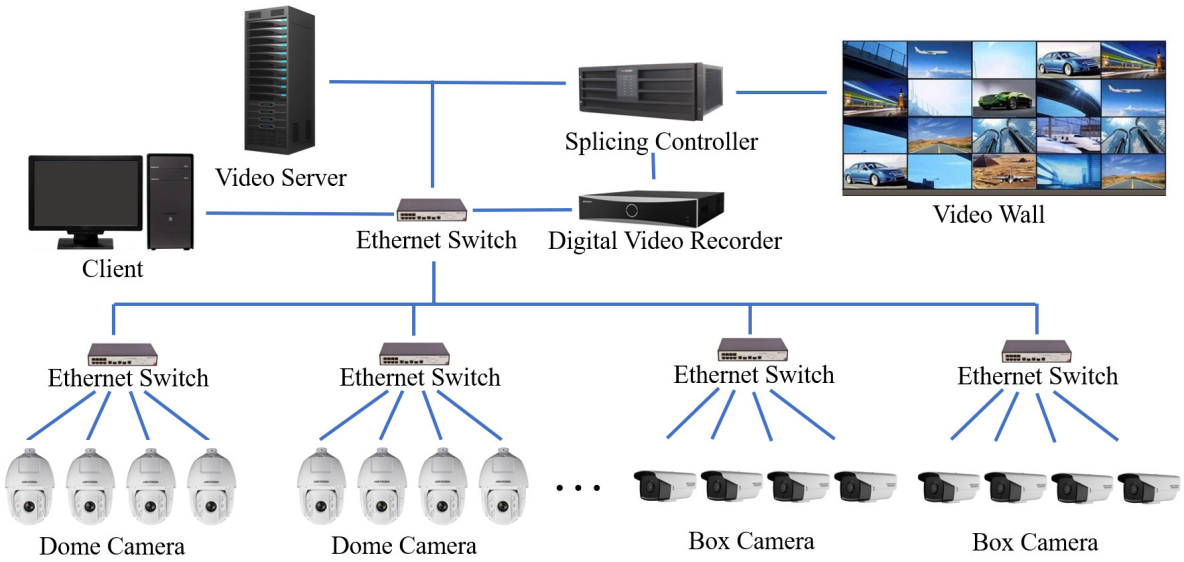


FIGURE 1. The Proposed Multi-Camera Video Surveillance System Framework

Some of them are correlated with each other. The latter satisfys the requirement of real time speeds. So we design our algorithm based on YOLOv3.

Another problem is abnormal behavior detection. The most common irregularities in port is the unsafe interactions between workers and vehicles. In the context of port, the majority of the cameras are palced in high positions. We regard pictures taken by these cameras as vertical views approximately, then measure the relative positional relationship between people and vehicles and evaluate the risk of the accident. We define a threat-based model inspired by [20] to assess risk. There are also some other works [21–26] related to object recognition and abnormal behaviour detection.

## 3. The Proposed System.

### 3.1. The Framework of System.

Our proposed intelligent multi-camera video surveillance system represented in Figure.1 integrates the pipeline of shooting, processing and visualization. All the cameras are connected to the connect the network video recorder through the ethernet switch. The splicing controller selects a certain number of video according to system setting or client command to display on the video wall. In this way, the watchkeeper can monitor the whole port. We build a video server to assist or replace the human in filtering out the frames with useful information such as abnormal behavior, irregularities and emergency situations. The server process the frames captured by all cameras with deep learning algorithm and sort them by abnormality, then schedule the screens of the top ranked cameras to display on the video wall.

### 3.2. The Object Detection Algorithm.

YOLOv3 is an object detector that uses features learned by a deep fully convolutional neural network to detect an object. It sees the entire picture when both training and testing so that it enables the end-to-end training and real-time speed. YOLOv3 divides the input image into a $S \times S$ grid. There are three different scales $13 \times 13, 26 \times 26$ and $52 \times 52$, which help YOLOv3 get better at detecting small objects. A grid cell is responsible for detecting the object whose center falls into the grid cell. YOLOv3 predicts bounding boxes using dimension clusters as anchor boxes.

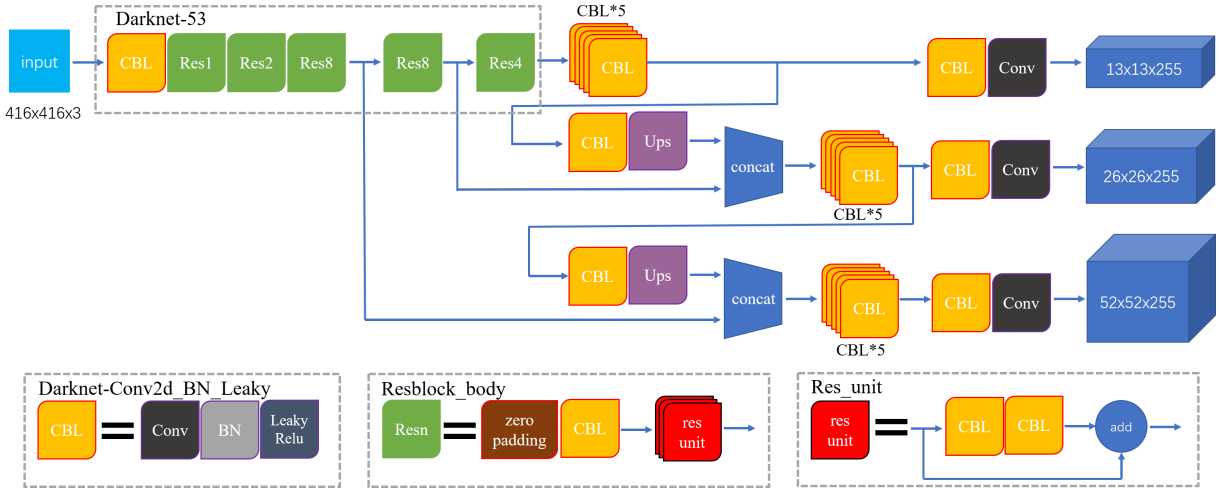


FIGURE 2. The Architecture of YOLOv3

The bounding box is described by 4 coordinates, $t_x$, $t_y$, $t_h$, $t_w$ such that

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. $(c_x, c_y)$ is offset from the top left corner of the image. $p_w$ and $p_h$ are anchors dimensions priors for the box. There are 9 bounding box priors determined by k-means clustering, 3 boxes are predicted at each scale. The output of each scale is a $S \times S \times [3 \times (4 + 1 + 80)]$ dimension tensor for the 4 bounding box offsets, 1 objectness prediction and 80 class predictions. Then the network performs non-maximum suppression to eliminate duplicate detections.The network for performing feature extraction is a hybrid approach with successive $3 \times 3$ and $1 \times 1$ convolutional layers as well as shortcut connections. The entire network structure show in Fig.2

3.3. **Threat-Based evaluation.** We introduce the notion of range of the closest point of approach(CPA) and time to the CPA. These two quantities are used for modeling two aspects of threat. We use this concept to assess the safety relationship between people and vehicle. The time and range to CPA for people $p$ and vehicle $v$ with corresponding state vectors $\mathbf{x}^{(p)} = [x^{(p)}\ v_x^{(p)}\ y^{(p)}\ v_y^{(p)}]^{\mathrm{T}}$ and $\mathbf{x}^{(v)} = [x^{(v)}\ v_x^{(v)}\ y^{(v)}\ v_y^{(v)}]^{\mathrm{T}}$ are given by

$$t_{CPA} = -\frac{pos_x vel_x + pos_y vel_y}{\sqrt{vel_x^2 + vel_y^2 + 0.01}} \tag{2}$$

$$d_{CPA} = \sqrt{(pos_x + t_{CPA} vel_(x))^2 + (pos_y + t_{CPA} vel_y)^2} \tag{3}$$

where

$$pos = [pos_x\ pos_y]^{\mathrm{T}} = [x^{(p)}\ y^{(p)}]^{\mathrm{T}} - [x^{(v)}\ y^{(v)}]^{\mathrm{T}} \tag{4}$$

$$vel = [vel_x\ vel_y]^{\mathrm{T}} = [v_x^{(p)}\ v_y^{(p)}]^{\mathrm{T}} - [v_x^{(v)}\ v_y^{(v)}]^{\mathrm{T}} \tag{5}$$

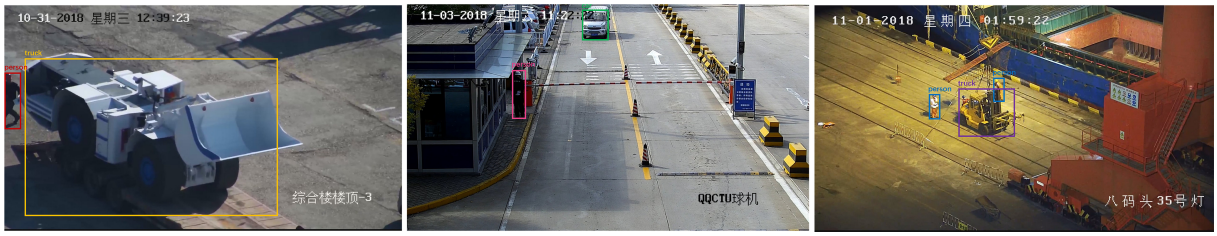


FIGURE 3. Qualitative Results

Then move from the time and range domain to the threat domain

$$\theta_t(\mathbf{x}^{(p)}, \mathbf{x}^{(v)}) = \begin{cases} 0, & -t_{CPA} < 0 \\ 1, & 0 \le -t_{CPA} \le t_0 \\ 1 - 2(\frac{t_{CPA}+t_0}{t_2-t_0})^2, & t_1 < -t_{CPA} \le t_1 \\ 2(\frac{t_{CPA}+t_2}{t_2-t_0})^2, & t_1 < -t_{CPA} \le t_2 \\ 0, & t_2 < -t_{CPA} \end{cases} \tag{6}$$

$$\theta_d(\mathbf{x}^{(p)}, \mathbf{x}^{(v)}) = \begin{cases} 1, & d_{CPA} \le d_0 \\ 1 - 2(\frac{d_{CPA}-d_0}{d_2-d_0})^2, & d_1 < d_{CPA} \le d_1 \\ 2(\frac{d_{CPA}-d_2}{d_2-d_0})^2, & d_1 < d_{CPA} \le d_2 \\ 0, & d_2 < d_{CPA} \end{cases} \tag{7}$$

where $t_0 < t_1 < t_2$ and $d_0 < d_1 < d_2$ are the points where the threat is equal to 1, 0.5 and 0. After mapping to the same domain, it is meaningful to aggregate them using a weighted sum to the total threat:

$$\theta(\mathbf{x}^{(p)}, \mathbf{x}^{(v)}) = \sum_{j=1}^{K} m_j \theta_j(\mathbf{x}^{(p)}, \mathbf{x}^{(v)}) \tag{8}$$

where $m_j$ is the weight of $\theta_j$ and $\sum_{j=1}^{M} m_j = 1$. For multi-target scenarios, the range and time to CPA are considered between each people and vehicle. For example, there are $n$ people and $m$ vehicles, the threat value $t$ can be selected such that

$$t = \arg\max \theta(\mathbf{x}^{(p_i)}, \mathbf{x}^{(v_j)}) \qquad i = 1, \ldots, n, j = 1, \ldots, m \tag{9}$$

The foundation for building this model is that we regard the picture taken at a high altitude as a top view of ground. The threat assessed not only based on the distance between people and vehicle, but also according to the direction of relative speed and relative motion. The same direction means that collisions will occur within a certain period of time. The length of time is used to define the value of threat. and the opposite direction means no collision. It will be demonstrate in the next section.

## 4. **Experiment.**

In the following section, we demonstrate that the proposed system can identify objects effectively and assess risk intelligently.

### 4.1. **Accuracy and real-time.**

We train the network at the full $416 \times 416$ resolution for 30 epochs on the COCO detection dataset. It detects an image in $29ms$ on a Nvidia 1080Ti GPU. It achieves good accuracy under real-time performance. Embedding our trained model into our system directly, it can still predict good bounding boxes and calssification in the live shot of the port. There are some results showing in Fig.3

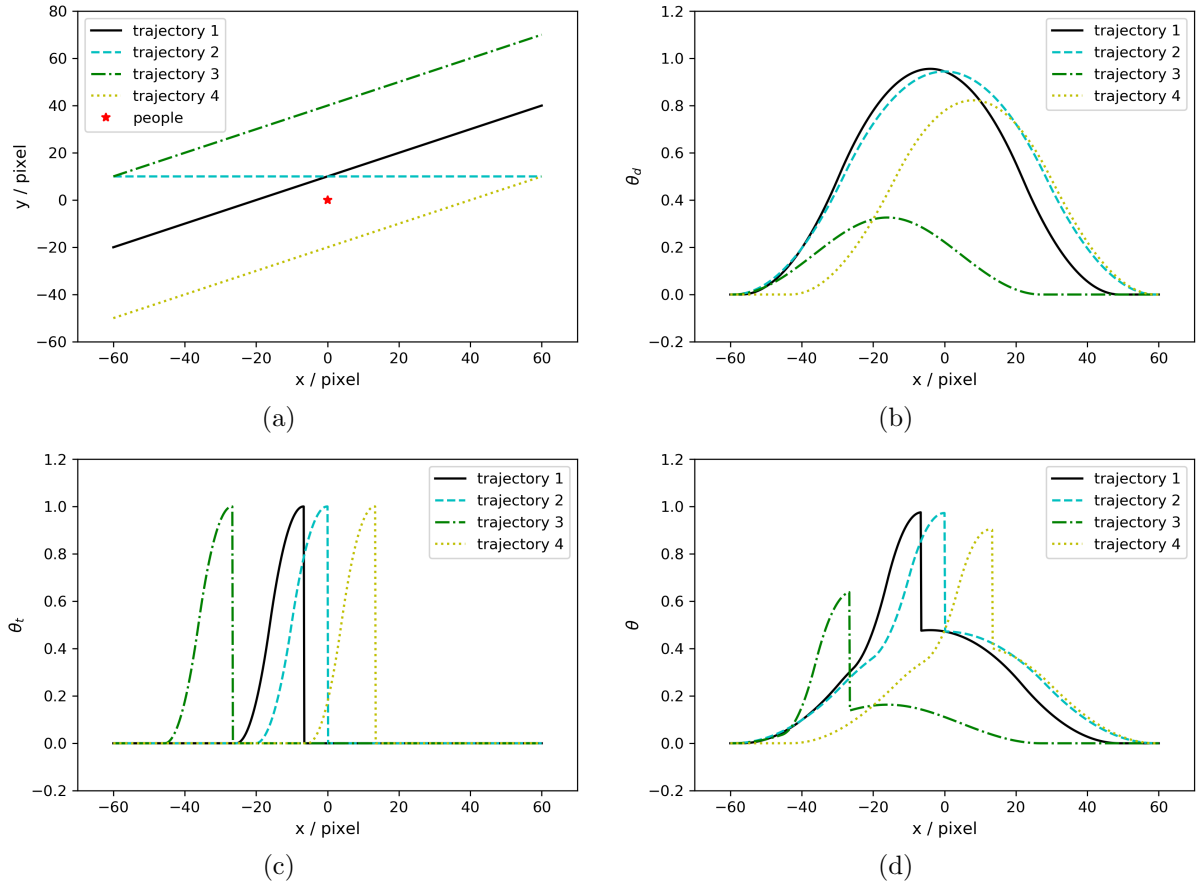


FIGURE 4. Threat evolution of considered trajectories

4.2. **Assess reasonablely.** The system gives meaningful alarms or push important pictures on the video wall. This is done by defining the threat model. In the example, the vehicles move along four trajectory in 2-dimensional space with a constant velocity. As shown in Fig.4, subgraph (a) demonstrates a geometry of example.Trajectory 1,3,4: The vehicles have constant velocity of [1,1] pixel/s. Trajectory 2: The vehicle has constant velocity of [1,0] pixel/s. Other parameters of threat model are $[t_0\ t_1\ t2] = [3\ 10\ 20]\ pixel/s$ , $[d_0\ d_1\ d_2] = [0\ 30\ 60]\ pixel$ and equal weights $m_1 = m_2 = 1/2$. Subgraph (b) shows the range threat of vehicle to people, the closer the distance, the greater the threat. Subgraph (c) shows the time threat to people. The shorter the time required for the vehicle to reach the position where people stand, the greater the threat. If the direction of displacement is opposite to the direction of speed, it means that people and vechile won't collide, so the time threat would be 0. Subgraph (d) shows the weighted threat. The advantage of this model is that reduce the effect of distance on the results. In other word, it is possible to have different threat values at the same distance, for example, approaching people and leaveing people. Obciously the vechile leaving people creates a smaller threat. Since the threat value ranges from 0 to 1, the system can sort cameras or videos based on threat. It is more efficient than manual selection when the number of cameras is large enough.

5. **Conclusion.** This paper proposes an efficient intelligent video surveillance system based on object recognition and abnormal behavior detection. We use YOLOv3 for real-time micro object detection, and threat-based method for assessing abnormal interactions between people and vehicles. The system changes the expensive and inefficient situation where human operators are assigned to monitor the video continuously to detect any

suspicious activity. Now the system is responsible for extracting crucial information or relevant scenes, and just presents the most necessary scenes to be monitored for operators.

Although the new system brings convenience to humans, it only analyzes the most common anomalies. Our model for object detection can identify many categories, and we haven't play the full potential of the model. In addition, we will collect dataset from the port, and train our model on it to improve recongnition accuracy. And properly reduce the network layer to becomes a lighter model with better performance in terms of speed.

**Acknowledgment.** This work is partially supported by the financial support from the Zhejiang Provincial Natural Science Foundation of China under grant No. LY16F010019, No. LY17F020019, No.19F030005, and the National Natural Science Foundation of China under grant No.61633019 and No.61972350. This research work is also partially supported by Ningbo Science and Technology Innovation 2025 major project(2021Z010,2019B10116 and 2018B10020), National Key R& D Program of China (No.2018YFB1702200) and SKLICT-OpenProjectProposal-2020(ICT20006)

## REFERENCES

[1] Y. Ye, S. Ci, A. K. Katsaggelos, Y. Liu, and Y. Qian, Wireless video surveillance: a survey, *IEEE Access*, 1: 646-660, 2013.

[2] N. N. A. Aziz, Y. M. Mustafah, A. W. Azman, A. A. Shafie, M. I. Yusoff, N. A. Zainuddin, and M. A. Rashidan, Features-based moving objects tracking for smart video surveillances: a review, *International Journal on Artificial Intelligence Tools*, 27(2):1830001, 2018.

[3] B. Sabri, Z. Ibrahim, M. M. Saad, N. N. A. Mangshor, and N. Jamil. Human detection in video surveillance using texture features, *The 6th IEEE International Conference on Control System, Computing and Engineering*, pp.45-50, 2016.

[4] K. R. Konda, Y. T. Tefera, N. Conci, and F. G. B. De Natale, Real time moving object detection and segmentation in H.264 video streams, *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp.1-6, 2017.

[5] M. Li, Z. Zhang, K. Huang, and T. Tan, Rapid and robust human detection and tracking based on omega-shape features,*The 16th IEEE International Conference on Image Processing*, pp. 2545-2548, 2009.

[6] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan, and M. Zaharadeen, Automated daily human activity recognition for video surveillance using neural network, *IEEE 4th International Conference on Smart Instrumentation, Measurement and Application*, pp. 1-5, 2017.

[7] M. Hussain, R. Xie, L. Zhang, M. Nawaz, and M. Asfandyar, Multi-target tracking identification system under multi-camera surveillance system,*International Conference on Progress in Informatics and Computing*, pp. 311-316, 2016.

[8] Z. Zhao, P. Zheng, S. Xu, and X. Wu, Object detection with deep learning: a review, *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212-3232, 2019.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation,*IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014.

[10] R. Girshick, Fast r-cnn, *IEEE International Conference on Computer Vision*, pp. 1440-1448, 2015.

[11] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: towards real-time object detection with Region Proposal Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137-1149, 2017.

[12] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 936-944, 2017.

[13] K. He, G. Gkioxari, P. Dollár and R. Girshick, Mask R-CNN, *IEEE International Conference on Computer Vision*, pp. 2980-2988, 2017.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You only look once: unified, real-time object detection, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[15] W. Liu,D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, SSD: Single shot multibox detector, *European Conference on Computer Vision*, pp. 21-37, 2016.

[16] J. Redmon, and A. Farhadi, YOLO9000: better, faster, stronger, IEEE Conference on Computer Vision and Pattern Recognition, pp. 6517-6525, 2017.
[17] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, DSSD: Deconvolutional single shot detector, *arXiv:1701.06659*, pp.1-12, 2017.
[18] Z. Shen, Z. Liu, J. Li, Y. Jiang, Y. Chen, and X. Xue, DSOD: Learning deeply supervised object detectors from scratch,*IEEE International Conference on Computer Vision*, pp. 1937-1945, 2017.
[19] J. Redmon, and A. Farhadi, YOLOv3: An incremental improvement, *arXiv:1804.02767*, 2018.
[20] S. J. Lee, and S. H. Jung, Threat-based sensor management for target tracking, *IEEE Transactions on Aerospace and Electronic Systems*, 51(4):2772-2785, 2015.
[21] K. Wang, C. M. Chen, M. S. Hossain, G. Muhammad, S. Kumar, and S. Kumari, Transfer reinforcement learning-based road object detection in next generation IoT domain, *Computer Networks*, 193: 108078, 2021.
[22] K. K. Tseng, J. Lin, C. M. Chen, M. M. Hassan, A fast instance segmentation with one- stage multi-task deep neural network for autonomous driving, *Computers and Electrical Engineering*, 93:107194, 2021.
[23] J. M. T. Wu, M.-H. Tsai, Y. Z. Huang, S. K. H. Islam, M. M. Hassan, A. Alelaiwi, and G. Fortino, Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model, *Applied Soft Computing*, 78:29-40, 2019.
[24] J. M. T. Wu, M. H. Tsai, S. H. Xiao, and Y. P. Liaw, A deep neural network electrocardiogram analysis framework for left ventricular hypertrophy prediction, *Journal of Ambient Intelligence and Humanized Computing*,online: 1-17, 2020.
[25] J. M. T. Wu, Z. Li, G. Srivastava, M. H. Tasi, and J. C. W. Lin, A graph-based convolutional neural network stock price prediction with leading indicators, *Software Practice and Experience*, 51(3):628-644, 2021.
[26] J. M. T. Wu, Z. Li, N. Herencsar, B. Vo, and J. C. W. Lin, A graph-based CNN-LSTM stock price prediction algorithm with leading indicators,*Multimedia Systems*, online, 2021.