Deep Learning and Sensor fusion technique for Indigenous Intelligent Vision system for Theft Detection

Pranav M, Veena N. Hegde*, Rachana S. Raju
Department of Electronics and Instrumentation Engineering,
B.M.S College of Engineering, Bangalore, India
pranavmahaveer.13@gmail.com,
veena.bms@gmail.com, rachanasraju@gmail.com

Roopa S.
Department of Electronics and Communication Engineering Siddaganga Institute of Technology,
Tumkur, India
roopas@sit.ac.in

2023 IEEE World Conference on Applied Intelligence and Computing (AIC) | 979-8-3503-1006-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIC57670.2023.10263839

Abstract— This paper proposes an indigenous system with modified architecture to classify different kinds of objects used in burglary and certain suspicious activities identified in thievery. The classifier used is a part of the security system to detect the images/ movements pertaining to theft or unusual acts while shoplifting valuable objects from jewellery stores. A sequence of integrated real-time computer vision algorithms and sensor fusion technology are implemented. The system adopts the transfer learning method (has 68 layers, 75 connections) as a pre-trained neural network to identify the unusual objects. Here, a person can take suitable action to aid the situation. The classifier algorithm was trained on a dataset of 144 cases of images and augmented appropriately belonging to seven different classes. It provides an accuracy of 97.73% during the validation phase and the confusion matrix resulted in an accuracy close to 95% and an F1-score of 84%.

Keywords— Computer vision, Indigenous system, Sensor fusion technology, Shoplifting objects, Theft identification.

I. INTRODUCTION

Business surveillance cameras prominently placed can aid to deter theft. A survey showed that 65 percent of all small businesses fall victim to employee theft, $45.2 billion in 2015 to retail theft, more than 1.38 percent of overall sales, and nationally, small businesses lose substantial amount of their business revenue every minute to shop lifters [1]. Another survey showed that 60% of burglaries are forcible entry, breaking in windows, picking off locks, etc., business owners pay directly and indirectly for vandalism. Smart cameras can not only detect motion, measure objects, read vehicle number plates but also can recognize human behaviours [2]. Conspicuously placed cameras have been proven to reduce threats of violence and vandalism at businesses dramatically. The algorithms are embedded in hardware with a programmable logic device or microcontroller and operate at a data rate of 20MHz per channel. The effective processing rate is 40MHz because each image processing logic device can process two channels of the image data [3]. The network video surveillance has been discussed in [3, 4]. Target tracking in a cluttered environment is one of the most challenging problems of video surveillance and monitoring systems. The concepts of human activities accompanied by the whole past events using inertial sensors, focused on smartphones is developed in [5].

The utilization of OpenPose [6] enables the extraction of human pose information, obtained through training on the COCO dataset to generate human pose features. However, when dealing with dynamic scenes such as explosions, the accuracy of capturing fast-moving sections is compromised due to limitations in frame-rate and processing power. To overcome these challenges, the integration of vision and video applications into embedded systems has emerged as a potential solution, offering remedies to the issues faced by video surveillance systems [7]. A comprehensive overview of object detection frameworks, including convolutional neural networks (CNN), as well as several datasets and evaluation metrics, is discussed in [8-10]. Furthermore, recent advancements in 3D object recognition using CNN, alongside the progress made in 2D, 2.5D, and 3D image analysis, are presented in [11]. Anomaly detection via a double fusion framework that combines appearance and motion features is discussed in [12].

Another significant application of computer vision is object recognition [13], which encompasses both specific, manufactured objects and the determination of generic object classes. Additionally, computer vision plays a crucial role in estimating the pose and spatial trajectory of objects over time. In [14], a smart monitoring system is introduced, which minimizes human intervention by performing checks on homes and work areas utilizing image processing methods to detect suspicious motion. Comparison of detection and recognition rates under different requirements, such as occlusion detection, side face detection, and facial exaggerated expression, is discussed in [15]. In order to recognize various objects, visual features are extracted, providing a semantic and robust representation. Notable visual features include the scale-invariant feature transform, Histogram of Oriented Gradients (HOG), and Haar-like

979-8-3503-1006-1/23/$31.00 ©2023 IEEE
DOI: 10.1109/AIC.2023.162

features. Object detection methods, such as the fusion of Haar-like feature and HOG features, are developed for edge detection in [16]. Theft detection and the motion of thieves using CCTV footages and tracking them using image processing techniques, without the use of sensors have been investigated in [17, 18]. This system focuses only on object detection. Six deep learning models are used for snatch theft detection in [19]. It can be observed from the survey that till now no attempt has been made to integrate computer vision algorithms and sensor fusion technology in detecting suspicious activities in thievery.

Presently, three state-of-the-art algorithms for object detection in images have gained prominence: faster region CNN, you only look once (YOLO), and single shot multibox detector (SSD). YOLO, specifically designed for real-time applications [20] operates differently from traditional classifiers as it serves as an object detector. Considering these developments, a sophisticated and reliable vision system to address the growing demands of businesses worldwide is proposed in this paper. The main contribution of the work is to analyse the theft problems from various perspectives and realizing how closed-circuit television's (CCTV) intermission acts in the shoplifting. The whole act of shoplifting and correlating it to the computer vision applications. Secondly, arriving at a possible understanding in realizing the solution with a mix of hardware and software with sensor fusion technology and deep learning concepts.

The outline of the paper is structured as follows. Section II provides the specifications of the proposed system. The system implementation is discussed in section III. In Section IV, the outcomes and findings of our system are presented and examined. Finally, section V provides the conclusion of the work.

II. PROPOSED SYSTEM

The proposed system is to develop and deploy an embedded system with an integrated sensor fusion technology and reduced algorithmic cum software resilience. The focus is to make the system analyse acquired data in real time and then immediately flag up salient features and unusual scenes. These flags give actionable insights and alert the user of the system. Thereby, buying him/her enough time to take appropriate response to facilitate the situation. The main objective of the work is to detect a person entering the shopping mart/retail shop with a theft tool or an unusual object like a hammer or gun, to detect unusual facial movement trying to assess the engagement of the store-keeper and to detect the action of theft (say the movement of the hands across the counter to shoplift or barge in the store to rob the place). In order to meet these objectives, a schematic has been implemented whose block diagram is depicted in Fig. 1. The hardware consists of the embedded system of a smart camera along with sensor elements. This embedded system is coupled to the video camera modules. These modules are high-definition image sensors. The video frames at a descent a frame rate is grabbed by the processor using these modules. The other sensor elements are those that add or validate information imparted by the main sensor, the camera.

The elements being a load cell is used for monitoring the total weight of all the items placed on a shelf or inside a counter. So, when a particular object is picked up from the shelf, the overall weight reduces which come in handy to validate the camera's recognition that an object has been picked up a person. Optic sensor/Proximity Infrared sensor (PIR) is used to monitor more or less the hand movement of the person picking the object from the shelf. Every time, the hand crosses the edge of the shelf, the sensor is activated and such information signal is passed on to the embedded system processor. PIR sensor is used for human motion detection; using this sensor the processor is able to keep a track of a person's movement within the store. To support the sensor interface, the embedded system consists of processor, memory unit, communication Interface and power supply unit. A Raspberry Pi 3 is used to just validate the algorithm and hence, the sensor fusion technology in the development stage. The communication interfaces Firewire, USB link, Gige, Camera Link, I2C are implemented to establish the communication between processor and different sensor modules. It is with this communication interface that the sensor elements mainly camera sensors, processor, memory unit and user interface are integrated into a smart camera system.

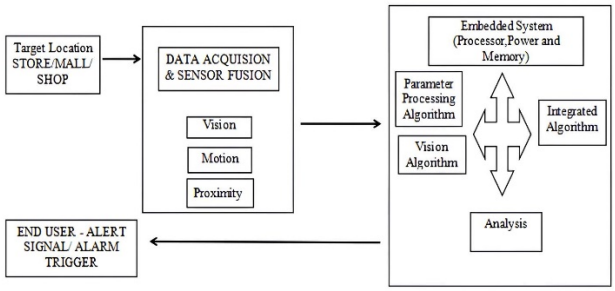


Fig. 1. Overall Block diagram of the proposed system.

The software involves the computer vision algorithms with a real time kernel, Raspbian OS. The algorithms used are intelligent pattern recognition algorithms, can detect motion, measure objects, recognise objects and faces, can track person and recognise human behaviours. Presently, the information received from the sensor elements and processed video frame information is further analysed to finally predict if a particular person has shoplifted the store. The user interface involves the transmission of the predicted information to the user. This information would be transmitted using wireless communication such as a wrist band alert or an alarm or a notification on the user's mobile phone thus, identifying the target shoplifter.

III. IMPLEMENTATION

The processor module gets the video input from the camera module. Further motion and optical sensors are connected to processor module. The algorithms developed using open CV on Raspberry Pi gets updated based on the results obtained and indicate these to processor module. The

steps followed for implementing the proposed system are described as follows.

A. *Feature Mapping and Bounding boxes-People and face detector*

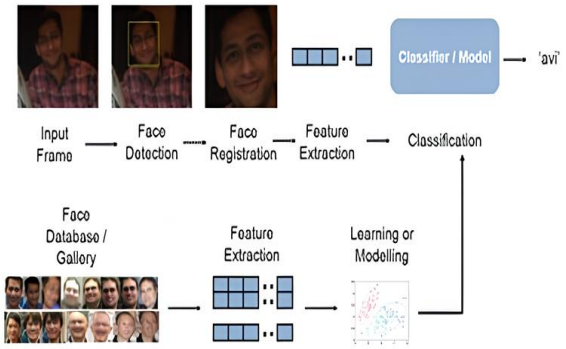


Fig. 2. Face Detector Algorithm - Process Flow.

People and face detector feature matching and bounding boxes algorithm is implemented using MATLAB and is shown in Fig. 2. Here person's facial image is read, HOG features are extracted, and support vector machine (SVM) classifier is used to detect person's face.

B. *Feature Extraction and Matching-Unusual object detector*

Feature extraction by identifying the feature points and matching the putative points is made to detect the object. YOLO operates differently from traditional classifiers as it serves as an object detector. YOLO divides the input image into a grid of S×S cells, where each cell predicts five bounding boxes that describe the object's position. Additionally, YOLO outputs a confidence score, which reflects the certainty of object enclosure. The score is solely related to the shape of the bounding box and does not provide information about the object's identity. YOLO algorithm for classifying objects is done using open CV. Deep learning algorithm like Squeezenet and Alexnet is used to classify objects. Face Tracking like KLT algorithm is used to identify facial features and track the points

C. *Motion Detection and Farneback method*

Surveillance Model such as sum of absolute differences (SAD) is used for motion detection. Simulink Model for theft video surveillance is shown in Fig. 3.

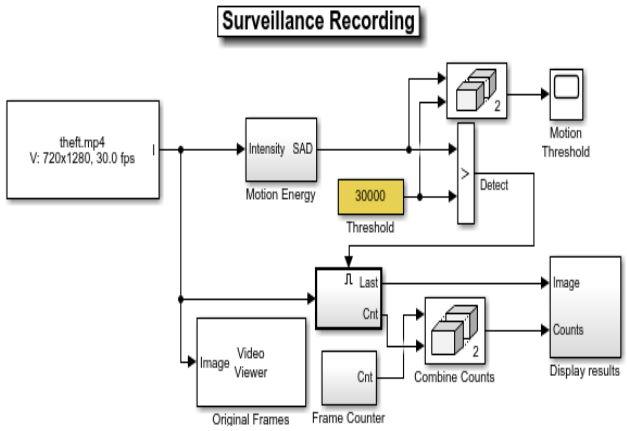


Fig. 3. Simulink Model for Surveillance.

Optical Flow - Farneback method is used for motion tracking. In this, video file is read and video frame is stored between two specific time intervals. The stored video object is converted to gray scale and calculated frame differences to remove stationary objects. Finally, flow on differenced frame is estimated and optical flow vectors on differenced frame is plotted.

IV. Results and Discussion

A. *The Algorithm*

The hardware components integrated together and is shown in Fig. 4. The entire process is executed in real-time and different conditions are checked for the following scenarios:

1) A person enters with arms (say, gun) to attack and loot the place.
2) A person enters casually, after sometime he/she observes that only the store keeper is present in the store, removes his/her knife to threaten and loot the place.
3) A group of people just barge in, create havoc and motion. They rob the place.
4) A person shoplifts just a few items by moving his/her hand across the counter, while the shopkeeper is being continuously engaged by someone else.

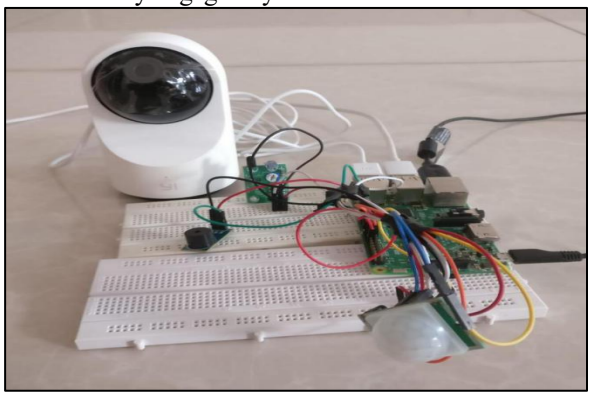


Fig. 4. Hardware prototype with all the components integrated together.

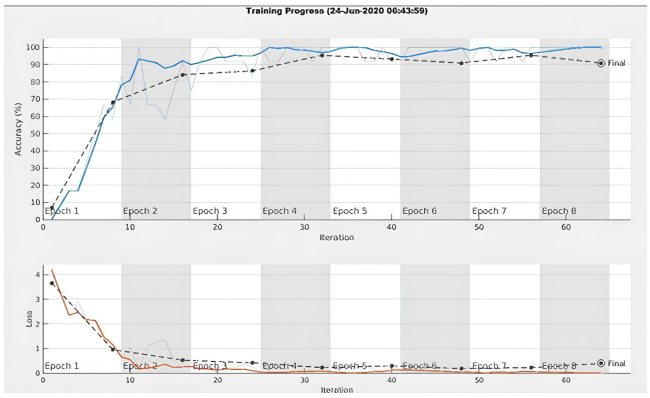
The hardware prototype is placed at the counter of a mimicked Jewellery store as shown in Fig. 5. The Dome X camera is kept at the far end of the counter. This captures live feed such as, movement of the hand across the counter in order to shoplift (scenario-4). Also, it can detect unusual objects such as a knife or gun that is intended to harm the store-keeper in order to loot the store (scenario-1 and 2). The PIR sensor is placed at a location that is closest to the valuable items out for display on the counter. Also, a series of proximity sensors can be embedded within the glass to track and verify the movement of hand across the counter (scenario-4). (Here, right to left movement by the shoplifter). This helps in reassuring and activation of the sensors, when the item is picked up and IR beam can't detect the item any longer.

Fig. 5. Hardware prototype (Distributed sensor arrangement) implemented at the counter section of mimicked jewellery store.

The motion sensor validates intense motion across the counter (suppose, the shoplifters jump across the counter and pounce on the store-keeper, as described in scenario - 3). The CCTV cameras (Here, two in number) are placed at appropriate angles that capture live feed of all activities within the store and across the counter. The sensor signals and the Dome X camera signal is sent to the Raspberry Pi 3. The embedded controller can be used as a standalone application by deploying the algorithm on it. The Raspberry Pi acts as a data acquisition device and output device. The data acquired is sent to a server which is then loaded and read on the online version of MATLAB (Cloud Computing). Even the live data feed of the CCTV cameras is uploaded to a server, which is then loaded and read on MATLAB. The data read along with real-time image classification algorithm saved as shopliftnet which is a transfer learnt and custom trained neural network along with people detector and face detector functions. The algorithm also checks for PIR and IR sensor data threshold crossings which then evaluates the SAD and optical flow algorithms, respectively. If the scenarios are validated, an alarm signal is generated at the output side of the embedded system using a buzzer.

B. Results

Training Dataset of 107 observation images and validation/Testing Dataset of 44 images are used. The CIFAR-10 dataset consists of 60,000 low-resolution images in 10 classes, commonly used for training the algorithm to identify weapons used in shoplifting. The dataset of images of the following seven classes are considered and identified as follows: Human as 'h', knife as 'k', pistol as 'p', scissors as 's', screwdriver as 'sd',weapon as 'w', wrench as 'wr'. The extracted data folder is selected and fed to Deep Network Designer (DND). Image Augmentation features are selected as random reflection in the x-axis and in the y-axis. The range from -90 to 90 degrees is chosen for random rotation. The range of 1 and 2 is taken for random rescaling. The amount of training data can be increased by applying randomized augmentation to data. Augmentation enables to train networks with less distortions in image data. Validation data is imported from the training data or from alternative source. Compared to the training data, model performance is estimated on new data, and safeguard against overfitting. Here, 30% of the images are used for validation. The pre-trained neural network is also analysed for its activations and learnable (Weights and biases) at every layer.



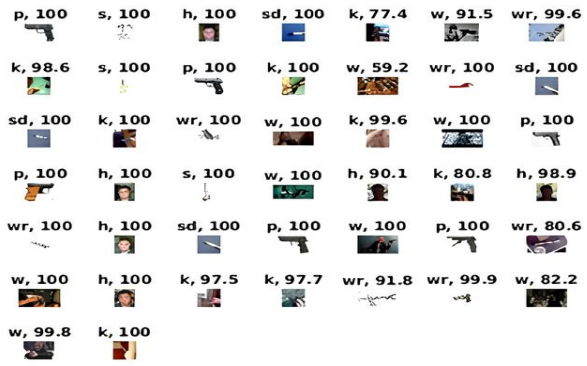


Fig. 6. Obtained Scores for the entire set of validation samples.

The required training options are selected as follows: The image classification networks have been trained using DND app. The network in DND is built for different types of data and then transfer the network for training. For large data sets, the default training options are better suited. The mini batch size and Validation frequency are reduced for small data sets. To obtain better control over the training, select the settings to train in training options. In order to slow down learning in the transferred layers, a small value is chosen in Initial learn rate. Validation Frequency is specified so as to calculate the correctness on the validation data once every epoch. For transfer learning, training for as many epochs are not needed. Many images are used by specifying the mini-batch size in each iteration. To make sure the entire data set is considered during each epoch, the mini-batch size is set so as to equally divide the number of training samples.

The solver used is "sgdm". The setting of various training option parameters are Initial Learn Rate to 0.0001, Validation Frequency to 5, and Max Epochs to 8. As there are 107 observations, Mini Batch Size is set to 15 to divide the training data evenly and use the whole data set during each epoch. For transfer learning of the network, the final learnable layer, classification layer and the output layer are replaced. A new classification Layer is placed onto the canvas in the layer library. The original output layer is removed and new layer is connected there. For a new output layer, data is not needed to set the output size. DND automatically fix the output classes of the layer from the data. Fig. 6 shows the training process for 8 epochs where accuracy and loss are plotted for many iterations and scores obtained for the validation samples. From the results, it can be observed that as the learning rate, maxEpochs, Weight Learn Rate Factor and Bias Learn Rate Factor increases the time elapsed and the overall accuracy increases.

A. *Sequence of analysis by the algorithm*

Scenario-1 : A person enters with arms to attack and loot the place. The algorithm initially read input video frames and resize the frame according to input layer size. People is detected and classified each frame using Squeezenet. Frame with desired label is matched out (Fig. 7) ex: weapon is identified. Matching based on pre-set accuracy threshold is done. Finally command send to raspberry pi to raise an alarm.

Fig. 7. Weapon identified in a scene.

Scenario – 2 : A person enters casually, after sometime he/she observes that only the store keeper is present in the store. Thief removes his/her knife to threaten and loot the place. It is identified by reading input video frames, face is detected and tracked the movements using KLT Algorithm (Fig. 8). Frame with desired label is matched out ex: knife is identified. Matching based on pre-set accuracy threshold. Command send to raspberry pi to raise an alarm.

Fig. 8. Knife identified in a scene.

Scenario - 3: A group of people just barge in, create havoc and motion. They later, rob the place. This is detected by reading input video frames, SAD algorithm identifies fast and intense movement. Motion sensor value is checked continuously for a period of time and activates as value crosses threshold for a set number of times. Unusual motion is detected by using frame 265. Later command send to raspberry pi to raise an alarm.

Scenario - 4 : A person shoplifts just a few items by moving his/her hand across the counter, while the shopkeeper is being continuously engaged by someone else. Hand movement is identified by a set of proximity sensors embedded under the glass counter (Left to right and re-tracking activation is checked - to conclude hand movement is not from behind the

counter). Input video frames are read, optical flow Farneback method algorithm confirms hand movement by tracking vectors and command is send to raspberry pi to raise an alarm.

TABLE I. MATRIX REPRESENTS THE OBTAINED ACCURACY, PRECISION, RECALL AND F-1 SCORE FOR THE MENTIONED CLASSES.

	Accuracy	Precision	Recall	F-1 Score
Knife	0.93	0.77	0.875	0.82
Scissors	1	1	1	1
Pistol	0.98	1	0.85	0.92
Screw-Driver	0.93	1	0.57	0.72
Human	0.95	0.83	0.83	0.83
Weapon	0.88	0.67	0.75	0.7
Wrench	0.95	0.71	1	0.83

TABLE II. OVERALL RESULTS OF THE CONFUSION MATRIX

Accuracy	0.95
Precision	0.85
Recall	0.84
F-1 Score	0.84

For accuracy validation, confusion matrix is used. For performance evaluation, multi-class classification algorithm is used with 44 number of samples. The classes are knife, scissors, pistol, screwdriver, human, weapon and wrench. Table I and II represents the matrix for obtained accuracy, precision, recall and F-1 score for the mentioned classes.

V. CONCLUSION

In this work, a system is proposed to address a reliable vision solution for businesses worldwide by using various computer vision techniques. The system developed is used in real-time with reduced resilience, complexity, processing power and computation time. The proposed system is an indigenous device that can cater to jewellery shops. Here a sequence of integrated real-time vision applications such as facial recognition and tracking, object detection and tracking, motion detection and tracking among others using embedded system hardware, sensor fusion technology and computer vision algorithms are implemented. The classifier algorithm was trained on a dataset of 144 cases of images and augmented appropriately belonging to seven different classes. It obtained an accuracy of 97.73% during the validation phase and the confusion matrix constructed resulted in an accuracy close to 95% and an F1-score of 84%. The system is also able to identify the other scenarios of robbery.

REFERENCES

[1] Cantrell, V. & Moraca, B. Organized Retail Crime Survey. Washington, DC: National Retail Federation. https://nrf.com/resources/retail-library/2015-organized-retail-crime-survey,2015.

[2] Yu Shi, Serge Lichman, "Smart Cameras: A Review", National Information and Communications Technology Australia (NICTA), Australian Technology Park, Bay 15 Locomotive Workshop, Eveleigh, NSW1430, Australia.

[3] W. Wolf, B. Ozer and T. Lv, "Smart Cameras as Embedded Systems", Computer,Vol.35,No.09,2002,pp.48-53,doi: 0.1109/MC.2002.1033027

[4] Chandrashekhar D. Badgujar, Dipali P.Sapkal, "A Survey on Object Detect, Track and Identify Using Video Surveillance", IOSR Journal of Engineering, Vol. 2, No. 10, October 2012, pp. 71-76.

[5] Wesllen Sousa Lima et.al. "Human Activity Recognition Using Inertial Sensors in a Smartphone: An Overview", Sensors, Vol. 19, No. 14, 3213, 2019. https://doi.org/10.3390/s19143213.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei and Yaser Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields ", IEEE transactions on pattern analysis and machine intelligence, Vol. 43, No. 1, 2019, pp. 172-186.

[7] Ying-li et.al. "IBM Smart Surveillance System (S3): Event Based Video Surveillance System with an Open and Extensible Framework", Machine Vision and Applications, Vol. 19, No. 6,2008, pp.315-327, DOI: 10.1007/s00138-008-0153-z.

[8] Zhong-Qiu Zhao et.al. "Object Detection with Deep Learning: A Review", IEEE Transactions on Neural Networks and Learning Systems, pp. 99:1-21, DOI:10.1109/TNNLS.2018.2876865.

[9] Xiao, Y., Tian, Z., Yu, J. et. al. "A review of object detection based on deep learning", Multimedia Tools and Applications, 23729–23791, 2020. https://doi.org/10.1007/s11042-020-08976-6.

[10] Ravpreet Kaur, Sarbjeet Singh, "A comprehensive review of object detection with deep learning", Digital Signal Processing Journal, Vol. 132,103812, 2022. https://doi.org/10.1016/j.dsp.2022.103812.

[11] Singh, R.D., Mittal, A. & Bhatia, R.K. "3D convolutional neural network for object recognition: a review", Multimedia Tools and Applications 78, 15951–15995 2019. https://doi.org/10.1007/s11042-018-6912-6.

[12] Dan Xu, Yan Yan, Elisa Ricci and Nicu Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion", Computer Vision and Image Understanding, Vol. 156, 2017, pp. 117-127.

[13] Juan Wu, Bo Peng, Zhenxiang Huang, Jietao Xie, "Research on Computer Vision-Based Object Detection and Classification", Computer and Computing Technologies in Agriculture VI, Vol. 392, ISBN : 978-3-642-36123-4,2013.

[14] Jain, S. Basantwani, O. Kazi and Y. Bang, "Smart surveillance monitoring system", International Conference on Data Management, Analytics and Innovation, Pune, India, pp. 269-273, 2017, doi: 10.1109/ICDMAI.2017.8073523.

[15] Di Lu, Limin Yan, "Face Detection and Recognition Algorithm in Digital Image Based on Computer Vision Sensor", Journal of Sensors, Vol. 2021, Article ID 4796768, 2021, 16 pages, https://doi.org/10.1155/2021/4796768.

[16] Xia, C., Sun, SF., Chen, P., Luo, H., Dong, FM. "Haar-Like and HOG Fusion Based Object Tracking", - Advances in Multimedia Information Processing PCM Lecture Notes in Computer Science, Vol. 8879. Springer, Cham. 2014. https://doi.org/10.1007/978-3-319-13168-9_18.

[17] M. S. Munagekar, "Smart Surveillance system for theft detection using image processing", International Research Journal of Engineering and Technology, Aug. 2018.

[18] R. Kakadiya et. al. "AI Based Automatic Robbery/Theft Detection using Smart Surveillance in Banks", 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 201-204.

[19] Zamri, N.F.M. et. al. "Snatch Theft Detection Using Deep Learning Models", In: Arai, K. (eds) Proceedings of the Future Technologies Conference (FTC), 2022. Lecture Notes in Networks and Systems, vol 559.

[20] Peiyuan Jiang et.al. "A Review of Yolo Algorithm Developments", Procedia Computer Science, Vol.199, pp. 1066-1073, 2022, https://doi.org/10.1016/j.procs.2022.01.135.