Machine Learning Based Performance Analysis of Video Object Detection and Classification Using Modified Yolov3 and Mobilenet Algorithm

[1]T Mohandoss and [2]J Rangaraj
[1] Department of ECE, Annamalai University, Chidambaram, Tamilnadu, India.
[2] Department of ECE, (Deputed to GCT Coimbatore), Annamalai University, Chidambaram, Tamilnadu, India.
[1]mohandosst@gmail.com, [2]jsrd_jsrd@yahoo.co.in

Correspondence should be addressed to T Mohandoss : mohandosst@gmail.com.

Article Info
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202303025
Received 15 December 2022; Revised from 02 April 2023; Accepted 10 May 2023.
Available online 05 July 2023.

Abstract – Detecting foreground objects in video is crucial in various machine vision applications and computerized video surveillance technologies. Object tracking and detection are essential in object identification, surveillance, and navigation approaches. Object detection is the technique of differentiating between background and foreground features in a photograph. Recent improvements in vision systems, including distributed smart cameras, have inspired researchers to develop enhanced machine vision applications for embedded systems. The efficiency of featured object detection algorithms declines as dynamic video data increases as contrasted to conventional object detection methods. Moving subjects that are blurred, fast-moving objects, backdrop occlusion, or dynamic background shifts within the foreground area of a video frame can all cause problems. These challenges result in insufficient prominence detection. This work develops a deep-learning model to overcome this issue. For object detection, a novel method utilizing YOLOv3 and MobileNet was built. First, rather than picking predefined feature maps in the conventional YOLOv3 architecture, the technique for determining feature maps in the MobileNet is optimized based on examining the receptive fields. This work focuses on three primary processes: object detection, recognition, and classification, to classify moving objects before shared features. Compared to existing algorithms, experimental findings on public datasets and our dataset reveal that the suggested approach achieves 99% correct classification accuracy for urban settings with moving objects. Experiments reveal that the suggested model beats existing cutting-edge models by speed and computation.

Keywords: Object Detection, Classification, Deep Learning, Image Classification.

I. INTRODUCTION

Understanding dynamic features in objects is critical in autonomous environments. The outdoor surveillance system employs freely moving event cameras. However, external variables make the structure not static, resulting in higher energy and time utilisation [1]. Using decades of machine vision research, we have addressed specialised object recognition challenges, including computerised assembly line sorting and inspection systems, handwriting detection on postal sorting machines, and ATM bill inspection. Despite these successful uses, the appearance of objects can be summarised in a well-controlled sensing environment, resulting in reliable and practical solutions for industrial difficulties that include perception and robot navigation [2].

The most crucial thing to bear in mind is that event cameras are not generating output pixel intensity levels but rather accurately time-stamped spikes, which are defined as events exhibiting a sufficient shift in pixel capturing intensity. In the end, event cameras use less transmission bandwidth and only use a few hundred. To summarise, event-based cameras adopt a distinct method of visual imaging by concentrating on low-latency and lightweight algorithms [3]. The reliability of the adaptive neuro fuzzy inference system (ANFI Stability) for categorising objects that move in a Street View application was examined. Neuro-fuzzy modelling combines the benefits of fuzzy logic and neural learning models, helping the framework defend actions based on object classification judgements [4-6].

Convolutional Neural Networks (CNNs) minimize the requirement for physical feature extraction in object classification utilization, removing previously determined image classification features. CNNs extract features from photos directly. Deep learning models are currently entirely accurate in various applications ranging from image processing to autonomous feature extraction. Deep CNN architectures employ complex models [7]. A larger picture of data collection is required for increased precision. To execute computer vision tasks, including object categorization, finding, recognition, and object tracking, CNNs require massive labelled data sets [8].

In dynamic object identification algorithms, static cameras are at the heart of cutting-edge techniques. This necessitates the employment of moving event cameras with dynamic object detection techniques. Despite the enormous benefits of event cameras, there yet needs to be a noticeable discrepancy in performance for different vision challenges among event camera techniques and frame-based algorithms. Additionally, frame-based detection increases hardware complexity, including the need for strong GPUs to efficiently retrain and build next-generation object identification frameworks [9-10]. This paper provides a straightforward and energy-efficient method for object recognition and classification compared to prior efforts. Lastly, the bounding box regressor should be trained. Selective search can produce region recommendations with a good recall, but the recovered suggested regions are time-consuming and tiresome. There are also some modifications to address the issue of erroneous placement. Many solutions to these difficulties have been presented [11-12].

The following is the contribution of this paper:

- Collecting diverse real-time video dataset and converting it into frames.
- A new object detection technique based on YOLOv3 and MobileNet is created. To begin, the feature map determination approach in the MobileNet backbone is optimized based on receptive field analysis.
- YOLOv3-MobleNet incorporates a Kalman filter to eliminate high-frequency noise elements for a smooth video frame to recognize numerous targets in a single video frame. The smoothed image is then utilized to identify objects in the background image using a background subtraction technique with a moving window.
- This work creates a simple and efficient tracking and detecting system for long-term event camera tracking and describes the several benefits of utilizing event cameras for object identification and tracking

The article is structured in the following order: the second section discusses previous work, the third section goes through the proposed approach, the fourth section goes over the results and discussion, and finally, the fifth section concludes the work.

II. LITERATURE SURVEY

Alexander Kugele et al. (2021) suggested a hybrid deep neural network complete training architecture for event-based pattern identification and object detection, with a spiking neural network (SNN) backbone for effective event-based feature extraction, followed by classic simulations. A neural network (ANN) is in charge of solving classification and detection tasks simultaneously. To do this, traditional back propagation is combined with agent gradient training to propagate gradients within SNN layers. Without additional conversion stages, hybrid SNN-ANNs can be trained to produce high-precision networks that are substantially faster in computation than their ANN predecessors [13].

Etienne Perrault et al. (2020) addressed all of these challenges in the context of event-based tasks for object detection. Initially, we made public an initial large-scale, high-resolution object detection dataset. More than 14 hours of 1-megapixel event camera footage in automotive scenarios are included, as well as 25 million high-frequency tagged auto, pedestrian, and two-wheeler bounding boxes. Additionally, they describe an innovative recursive framework for event-based detection and time consistency loss in order to improve behavior training. Experiments on the data set described in this paper, which includes grayscale events and images, indicate comparable performance to highly tuned and widely researched frame-based detectors [14].

Bharath Ramesh et al. (2020) developed an event-based feature extraction approach by aggregating local activities in image frames and using principle component analysis (PCA) to normalize neighboring regions. As a result, the proposed system can be implemented in an FPGA tool, resulting in an excellent performance-to-energy ratio. The suggested approach outperforms current methods for object detection when evaluated on a data set based on real events [15].

Shixiong Zhang et al. (2022) developed an event camera based dynamic object-tracking system to accomplish long-period steady event object detection. Using an adaptive method to match the spatiotemporal scope of event data is a crucial innovative element of our approach. To that end, we use online learning to rebuild event images from rapid access to asynchronous streaming data. Unlike standard object tracking jobs that use a fixed camera, all three tracking scenarios include the camera and object violently rotating and shaking simultaneously. Experimental findings reveal that the suggested method surpasses previous state-of-the-art techniques regarding precision and resilience [16].

Miguel Angel et al. (2021) presented an event-by-event analyzing approach for employing UAS to identify human infiltration. These include: 1) recognizing clusters of events created by objects that move on a static background; and 2) calculating the chance that a group corresponds to a person using convolutional neural networks. The proposed technique has been implemented and tested in difficult conditions. The proposed technique was constructed and empirically validated in challenging, cluttered scenarios with varying illumination conditions and item types. The performance validation reveals

precision rates of more than 90%, accuracy values greater than 70%, and recall levels greater than 70%, verifying its intended functionality [17].

Srinivas et al. (2022) examined frameworks for these activities using better cyber security control facilities. The technique is divided into two stages: detection of numerous objects using the cyber security probabilistic Gaussian mixture model and background suppression, and tracking of multiple moving objects using the kernel convolution moving window with kalman filter. Simulations outcomes show that the suggested approach can identify and locate objects in complicated and shifting environments with excellent efficiency, resilience, and accuracy. This proposed model also yields a noise-free image [18].

Kyung Pyo Kim et al. (2020) proposed a methodology for enhancing deep learning (DL)-based identification accuracy using shape data gathered from LiDAR point clouds. This research also presents a layer-based building technique that takes into account the three degrees-of-freedom motion of dynamic objects in order to augment this shape information properly. In experiments, the suggested cumulative technique beats existing log-based algorithms. Furthermore, in the actual car data test, the DL algorithm trained on simulated data performed better when gathering the lidar point cloud [19].

Guray Sonugur et al. (2022) suggested a two-stage Interconnected Artificial Neural Network (ICANN) framework. At the end of the GPS-assisted picture registration procedure, live images are transformed into binary images in the first stage. The shape of a silhouette is then created by labelling related components in the image background. Two interlinked neural networks are employed in the second stage. The initial neural network determines if the outlines are objects or noise. The maximum success percentage for object classification in experimental investigations was 96.1%. The acquired findings are compared to the currently popular YOLO object identification technique [20].

OA Pakhomova et al. (2019) developed a method for implementing a motion-detecting approach to enhance the effectiveness of the movement vector search technique used by the detection subsystem; the basic idea is to break each frame into blocks and look for similar sections in the subsequent frames. The study outcomes demonstrate the method's efficacy. To remove them, a motion detection module is suggested to be integrated into a multipurpose machine vision framework that collects images from cameras at the input and communicates accumulated information on the objects seen through parallel streams at the output.

The detection module is in charge of searching for and detecting movement, as well as concealing extraneous information and presenting only the areas required for further classification [21].

TJing Yunduo et al. (2021) suggested an event camera corner extraction and tracking technique that is asynchronous in real-time. The primary motivation for this paper is to increase corner identification and tracking accuracy while maintaining computing efficiency. Lastly, to enable corner event tracking, we offer a data association strategy with temporal, velocity, and spatial direction constraints, in which they associate a recently arrived corner event with the last active corner in its neighbourhood that fulfills the speed direction requirement. The trials are carried out on the conventional event camera dataset, and the findings reveal that the technique performs exceptionally well in corner detection and tracking [22].

Justas Furmonas et al. (2022) summarize the approaches and systems based on events that have been reported and are now known. An examination of these approaches and frameworks analytically supports the findings reached. The paper finishes with suggestions and proposals for future improvements in the domain of events using chamber depth estimation. A recent study demonstrates the use of SNNs, unsupervised and supervised neural networks. Nevertheless, many approaches continue to perform poorly due to a shortage of suitable training data sets [23].

Takehiro Ozawa et al. (2022) suggested a method for predicting motion in bird's-eye view space using contrast optimization. This paper reduces the dimensionality to a 2D motion estimate rather than a 3D motion estimate by translating the dataset to a bird's-eye view employing homograph derived from the camera position. This conversion solves the issue of non-convex loss functions in previous approaches. The experimental findings with CARLA and real-world data show that the suggested approach is efficient and accurate [24].

Problem statement

A deeper design gives tenfold greater effusive capability when compared to standard shallow models. To achieve high detection accuracy, the following issues must be resolved:

- Intra-class variances: shape, size, material, colour, and position differences in real-world objects.
- Image circumstances and unconstrained surroundings: variables including blur, lighting, shadow, weather conditions, clutter, occlusion, physical object location, motion, and viewpoint.
- Imaging noise: compression noise, filter distortions, and low-resolution images are instances of imaging noise.
- The detector must discriminate between thousands of organised and unstructured real-world item categories.
- Low-end mobile devices possess restricted speed, memory, and processing capabilities.
- There should be distinctions between thousands of open-world object classes.
- Image or video data on a large scale.
- Impossibility of handling previously unseen objects.

The fundamental idea is to use a CNN on the image to complete the task. CNN performs tasks on image patches, and many of these highlighted regions can be produced utilizing region-suggested networks, including the Regional Convolutional Neural Network (RCNN), the Fast-Region Convolutional Neural Network (Fast-RCNN), and the Faster-Region Convolutional Neural Network (Fast-RCNN). A hierarchical clustering approach is utilized to do a selective object recognition search. These approaches have a few bottlenecks that can be addressed with cutting-edge techniques, including You Only Look Once (YOLO) and Single Shot Detector (SSD). An effective object identification method is a technique that recognizes bounding boxes for all real-size objects while using powerful computing resources and a faster processing speed. YOLO and SSD provide promising outcomes. However, there is a trade-off between speed and precision. As a result, the choice of method is application-specific [25].

In instances of dynamic objects in the background, the LIBS approach does not deliver the most accurate results. Suppose there is a slight change in the background, such as swinging a sheet or any other minor alteration. Only upright humans can be spotted in W4 utilizing the cardboard design. It becomes problematic when people are in various positions, crawling and climbing. The detection of spatial irregularities that include U-turns in behavioral subtraction is difficult in this technique. The study can identify both temporal and spatial outliers when need to identify them. When foreground objects become visible during background activities, behavioral camouflage occurs. The Kalman Filter, Mean Shift Algorithm, and GMM all struggle to detect multiple objects with minor occlusions. Conventional object detection techniques cannot identify areas in images with numerous objects. Existing color detection algorithms can only detect primary colors accurately.

Existing approaches detect colors incorrectly if the image contains other colors. Aside from that, some common difficulties are that if the background illumination changes, it could be misinterpreted as a front object. Some approaches also have difficulties in detecting shadows. The closeness in glimpses between foreground and background items can be problematic for camouflage. Another problem is non-static background modelling. In high-traffic locations, the background is frequently obscured by many foreground objects. Because of the constant shift, it makes it challenging to classify the permanent foreground and backdrop [26].

III. PROPOSED METHODOLOGY

This work aims to create feature selection, and classification approaches to address existing issues with the detection of moving objects collected using event cameras. Low-range approximation techniques are utilised to extract the dynamic properties of the frames. To minimise battery usage, a freshly enhanced YOLOv3 is employed for feature selection. The suggested approach assesses the frame's entropy, lowering power usage.

Additionally, to reduce the computational time consumption of the suggested enhanced YOLOv3, data set classification is accomplished by utilising YOLOv3 and MobileNet architecture. The ranking is accomplished through a comparative examination of live data sets. **Fig 1** depicts the basic framework for identifying targets.

MOT20 is a real-time video dataset acquired on this work, and it has been converted to tiny video frames. The suggested design and noise identification of several moving objects will be displayed in the discovered object's frame. A convolutional moving window Kalman filter is used to remove and smooth noise. The video frames will be processed and analyzed after denoising. Utilizing noisy measurements acquired over time, the Kalman filter estimates method parameters and predicts future observations. At every stage, it makes predictions, collects measurements, and subsequently updates based on the forecasts and comparisons. The mathematical estimator can predict and update the state of a wide range of linear processes. In the YOLOv3 network, the binary cross-entropy loss is utilized rather than multiple labels to classify for predicting the classes of bounding boxes to improve performance.

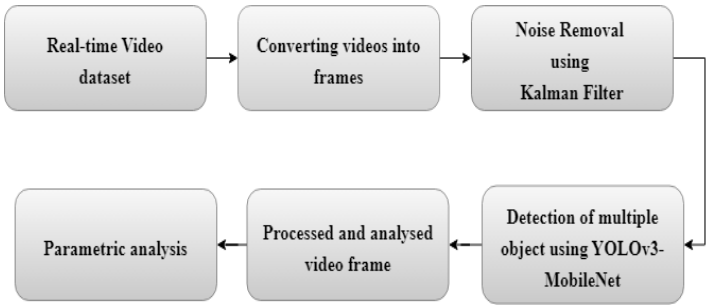

Fig 1. Basic Block Diagram of Object Detection

MobileNet

The MobileNet model is the backbone of the object detection architecture in this work because it is small and complex. Regular convolutions are divided into depth and point convolutions in the MobileNet model. Depth convolution divides

traditional convolution into two distinct layers for merging and filtering. The point convolution then uses a 1x1 convolution to mix the outputs of the deep convolutions [27]. **Table 1** show This factorization considerably reduces both the computation time and the size of the model. The calculation cost is computed utilizing the following operations:

$$d_k . d_k . m . d_f . d_f + m . n . d_f . d_f \quad (1)$$

Where m and n are the numbers of input and output channels, d_k denotes the convolution operation kernel size and d_f denotes the size of the feature map, respectively. The depthwise and pointwise convolution is followed by BN and ReLU blocks are shown in **Fig 2**.

The computing cost for standard convolution, on the other hand, is:

$$d_k . d_k . m . n . d_f . d_f \quad (2)$$

Combining (1) and (2) yields the following calculation reduction:

$$\frac{d_k . d_k . m . d_f . d_f + m . n . d_f . d_f}{d_k . d_k . m . n . d_f . d_f} = \frac{1}{n} + \frac{1}{d_k^2} \quad (3)$$

Table 1. Mobilenet Model Layer Architecture

Name	**Shape of the filter**	**Size of the input**
Conv	3 x 3 x 3 x 32	416 x 416 x 3
Conv dw	3 x 3 x 32 dw	208 x 208 x 32
Conv	1 x 1 x 32 x 64	208 x 208 x 32
Conv dw	3 x 3 x 64 dw	208 x 208 x 64
Conv	1 x 1 x 64 x 128	104 x 104 x 64
Conv dw	3 x 3 x 128 dw	104 x 104 x 128
Conv	1 x 1 x 64 x 128	104 x 104 x 128
Conv dw	3 x 3 x 128 dw	104 x 104 x 128
Conv	1 x 1 x 128 x 256	52 x 52 x 128
Conv dw	3 x 3 x 256 dw	52 x 52 x 256
Conv	1 x 1 x 128 x 256	52 x 52 x 256
Conv dw	3 x 3 x 256 dw	52 x 52 x 256
Conv	1 x 1 x 256 x 512	26 x 26 x 256
5 x Conv dw	3 x 3 x 512 dw	26 x 26 x 512
Conv	1 x 1 x 512 x 512	26 x 26 x 512
Conv dw	3 x 3 x 512 dw	26 x 26 x 512
Conv	1 x 1 x 512 x 1024	13 x 13 x 512
Conv dw	3 x 3 x 1024 dw	13 x 13 x 1024
Conv	1 x 1 x 1024 x 512	13 x 13 x 1024
Avg Pool	Pool 7 x 7	13 x 13 x 1024
FC	1024 x 1000	1 x 1 x 1024
Softmax	Classifier	1 x 1 x 1000

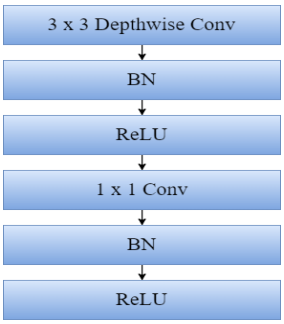

Fig 2. The depth-wise and point-wise convolution is followed by BN and ReLU blocks

Enhanced You Only Look Once v3 (YOLOv3)

In the conventional YOLOv3 framework, the Darknet-53 network serves as the basis of feature extraction. This network outperforms ResNet-152 and ResNet-101 in terms of power and efficiency. Nonetheless, Darknet-53 contains numerous layers, making it challenging to execute on mobile devices. Google's MobileNet framework [28] decomposes conventional convolutions into deep convolutions and 1x1 convolutions to minimize model size. It has been found that the MobileNet framework consumes 8-9 times less computing than existing convolutions, with a minor loss of precision. As a result, the MobileNet network rather than the Darknet-53 model serves as the basis of the YOLOv3 framework for object detection in this work.

The YOLO network's convolutional layers are inextricably linked to the underlying Darknet technology. Furthermore, for a more cohesive grid, they can be replaced with their pointy equivalents. The suggested object detection technique is provided in this section. A new object detection system is created using YOLOv3 and MobileNet**. Fig 3** depicts the suggested architecture. The suggested approach begins by rescaling image data from event based real time dataset. MobileNet is an essential feature extraction component in this approach due to its excellent accuracy and effectiveness. In contrast to the conventional YOLOv3 model's selection of fixed feature maps, this research reexamines how to determine the object detection feature maps using matching receptive field and object scale. The revised selection of feature map significantly improves the suggested object detection model's performance [29-30].

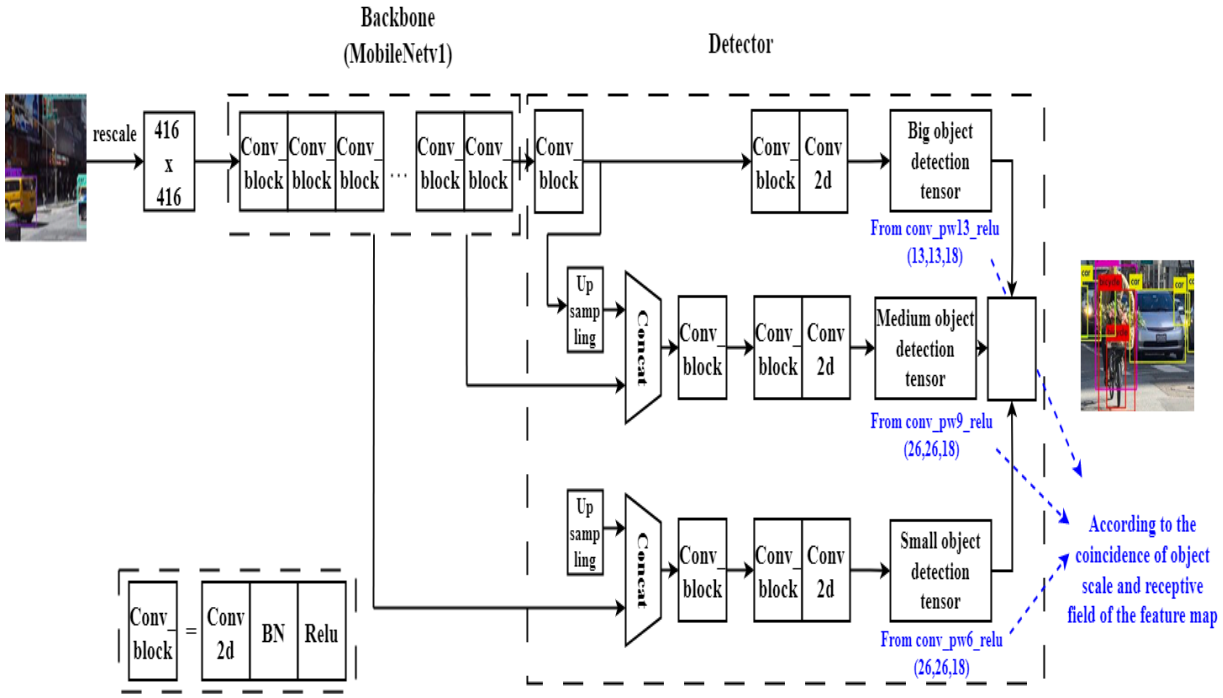

Fig 3. Proposed Method's Framework

Algorithm 1 Proposed object Detection Algorithm

Input : original image
Output : image with rectangle indicators
for i ← 1 to the number of scales in the image pyramid **do**
  Downsample image to generate $image_i$
  Calculate integral image, $image_{ii}$
  for j ← 1 to number of steps of sub-windows **do**
    for k ← 1 to number of stages is cascade classifiers **do**
      for l ← 1 to number of filters of stage k **do**
        Compile filter outputs
      end for
      if compilation fails per-stage threshold **then**
        Restrict the use of sub-window
        Break this k for loop
      end if
    end for
    if all per-stage checks were passed by the sub-window **then**
      Consider this sub-window an object.
    end if
  end for
end for

IV. RESULTS AND DISCUSSION

The simulation tool in the suggested strategy is MATLAB. MOT20 [31] is a real-time video dataset that can be used in this work. **Fig 4** depicts the results of the proposed YOLOv3-MobileNet processing. Precision, Accuracy, True Positive Rate (TP), False Positive Rate (FP), Ground Truth (GT), Mean Absolute Position (MAP), and Detection Rate (DET) are the parameters for the analysis.

Fig 4 . Result of the proposed model from the MOT 20-01 sequence of the real-time MOT 20 dataset.

Accuracy Analysis

The degree of agreement between an actual value and its noise evaluation is referred to as accuracy. Table 2 depicts an accuracy study of the proposed approach.

Table 2. Evaluation of Accuracy

Number of images	SSD	RCNN	F-RCNN	YOLOv2	YOLOv3	Proposed
100	77	82	90	91	93	95
200	81	85	92	93	95	97
300	83	87	93	95	96	98
400	85	91	95	96	97	98.5
500	87	93	96	97	98	99

On the X axis of **Fig 5**, multiple video frame sequences from the MOT 20 data sets are given, and the accuracy in percentage is assessed on the Y axis. This implies a maximum accuracy of 99%. The proposed model's accuracy estimations are validated against existing models using feature masking in video frames. The comparative study takes into account the MOT of ten items as well as the classification accuracy of the suggested model.

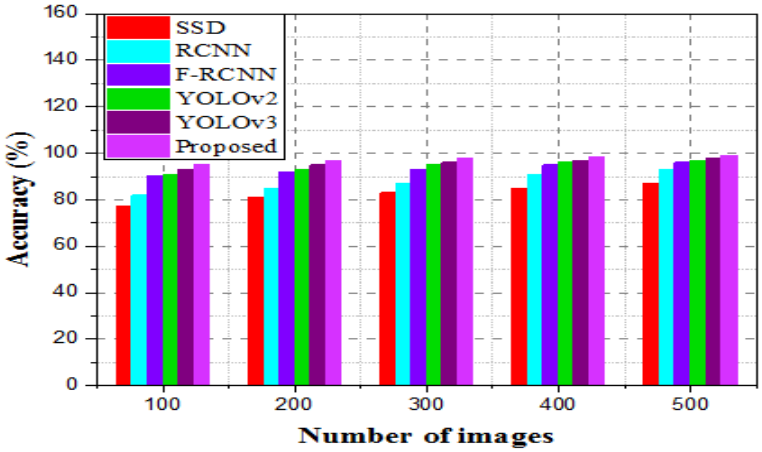

Fig 5. Analysis of Accuracy

Precision Analysis

Precision is the degree to which reiterated noise measurements produce the same results under similar circumstances. **Table 3** depicts the precision analysis of the suggested technique.

Table 3. Evaluation of Precision

Number of images	SSD	RCNN	F-RCNN	YOLOv2	YOLOv3	Proposed
100	73	77	82	85	88	90
200	75	78	86	87	89	93
300	77	79	88	90	93	95
400	79	80	89	92	94	96
500	81	82	91	93	95	98

On the X axis of **Fig 6,** multiple video frame sequences from the MOT 20 data sets are shown, while the precision in percentage is assessed on the Y axis. As a result, the proposed approach obtains the most excellent precision of 98%. The calculation of precision values reveals that the suggested model achieves precision levels that are superior to the state of the art. Existing technology provides accuracy rates of 95%, 93%, 91%, 82%, and 81% for SSD, RCNN, F-RCNN, YOLOv2, and YOLOv3, respectively. In the instance of the suggested model, the observed precision is 98%. Comparisons show that it outperforms traditional technologies.

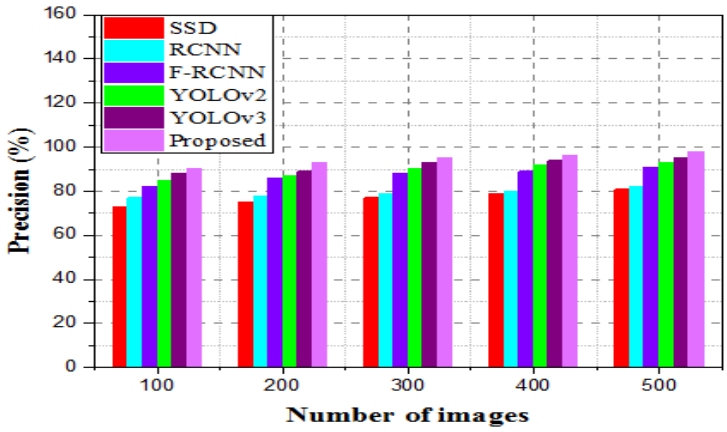

Fig 6. Analysis of Precision

Recall Analysis

Recall refers to the proportion of relevant images obtained overall. **Table 4** shows the suggested technique's recall analysis.

Table 4. Evaluation Of Recall

Number of images	SSD	RCNN	F-RCNN	YOLOv2	YOLOv3	Proposed
100	63	70	75	77	79	85
200	65	71	77	79	81	87
300	67	73	80	81	83	90
400	68	75	82	83	85	93
500	70	77	83	85	87	95

Fig 7 shows multiple video frame sequences from the MOT 20 data set on the X-axis and recovery percentages on the Y-axis. This proposed approach has a maximum recovery value of 95%. A comparison of the suggested method and the current state of the art reveals that the suggested approach outperforms conventional procedures. Model tracking and categorization outperform existing methods.

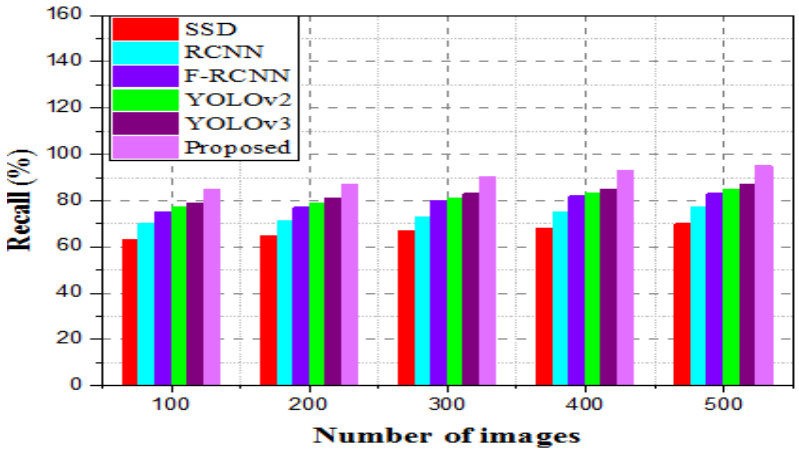

Fig 7. Analysis of Precision

TP Analysis

The criteria needed for evaluating the performance of a tracker is defined as True Positive analysis. The first step is to determine whether each proposed output is a TP that corresponds to an actual goal. The TP of the proposed approach is evaluated in **Table 5.**

Table 5. Evaluation of TP

Number of images	SSD	RCNN	F-RCNN	YOLOv2	YOLOv3	Proposed
100	53	55	59	61	64	78
200	55	57	61	64	66	84
300	57	61	65	66	68	88
400	58	63	67	68	70	90
500	61	64	68	70	72	95

Fig 8 shows multiple video frame sequences from the MOT 20 data set on the X-axis and true positive on the Y-axis. The suggested approach yields the most excellent True Positive value of 95%.

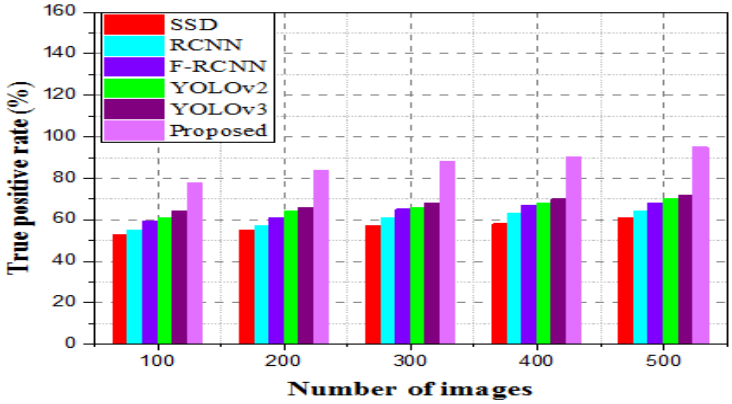

Fig 8. Analysis of TP

FP Analysis

The first step is determining whether each hypothesized output is an FP or false alarm. The study of false positives is shown in Table 6. FP can represent the number of images recognized or classified by a model per second in image classification and object detection applications. It can be utilized for estimating the model's average processing speed. The X-axis in **Fig 9** shows distinct video frame sequences from the MOT 20 dataset, while the Y-axis shows false positives. This proposed approach yields the most excellent false positive rate of 94%. The image field is defined by the FP value, which refers to the number of frames transmitted by the screen every second.

Table 6. Evaluation of FP

Number of images	SSD	RCNN	F-RCNN	YOLOv2	YOLOv3	Proposed
100	53	57	62	64	66	74
200	55	60	65	67	70	80
300	57	63	67	69	72	85
400	60	65	71	73	75	93
500	62	67	73	75	79	94

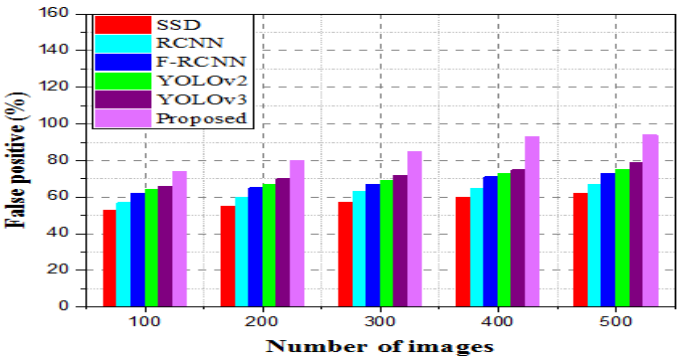

Fig 9. Analysis of FP

Ground Truth Analysis

The phrase "ground truth" implies the information collected in the field. Image data may be associated with real-world features and real-world material on the ground can be exploited. It is also helpful for atmospheric adjustment. Table 7 shows an investigation of the GT of the proposed method's actuality. On the X-axis of **Fig 10**, several video frame sequences from the MOT 20 dataset are shown, while the ground truth is considered on the Y-axis. The proposed work discovers the greater truth of this statement.

Table 7. Evaluation of Ground Truth

Number of images	SSD	RCNN	F-RCNN	YOLOv2	YOLOv3	Proposed
100	67	73	77	80	82	86
200	69	77	79	82	84	88
300	71	81	84	86	88	90
400	73	83	86	87	89	94
500	75	85	88	90	92	96

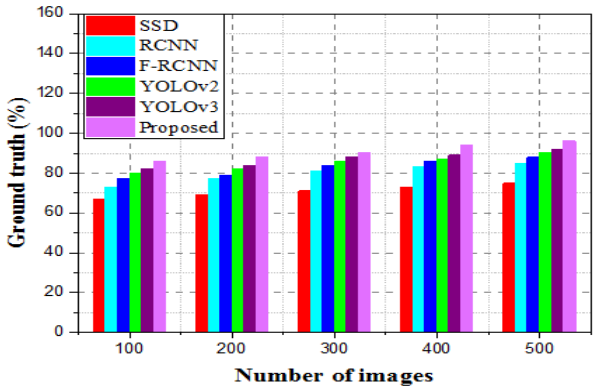

Fig 10. Analysis of Ground Truth

Detection Analysis

This analysis solely considers pedestrians. People who are static and other groups are shielded from detection and ground truth. Table 8 displays the assay results**. Fig 11** depicts several video frame sequences from the MOT 20 data set on the X-axis and detections on the Y-axis. The proposed technique surpasses all others in terms of detection value. This proposed approach yields the most excellent detection rate of 93%.

Table 8 Evaluation of Detection Rate

Number of images	SSD	RCNN	F-RCNN	YOLOv2	YOLOv3	Proposed
100	47	54	57	63	69	72
200	51	57	59	65	72	82
300	53	61	61	68	75	85
400	55	63	64	72	77	88
500	57	65	67	75	80	93

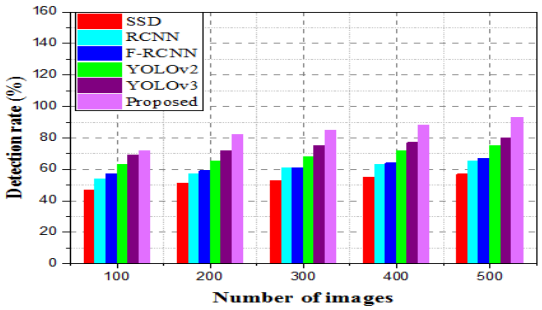

Fig 11. Analysis of Detection Rate

MAP analysis

The Position information containing coordinate details linked with a single image is called mean absolute position. The mean absolute position analysis in **Fig 12** is shown in Table 9, where several video frame sequences from the MOT 20 data sets are shown on the X-axis, and the MAP is examined on the Y-axis. The most excellent mean absolute position score obtained was 95%. A comparison of the suggested and state-of-the-art MAP scores reveals that the suggested approach outperforms the existing literature.

Table 9. Evaluation of Mean Absolute Position

Number of images	SSD	RCNN	F-RCNN	YOLOv2	YOLOv3	Proposed
100	57	63	67	71	74	75
200	61	65	73	75	77	79
300	64	67	75	77	79	83
400	67	70	77	81	83	87
500	71	74	81	85	87	95

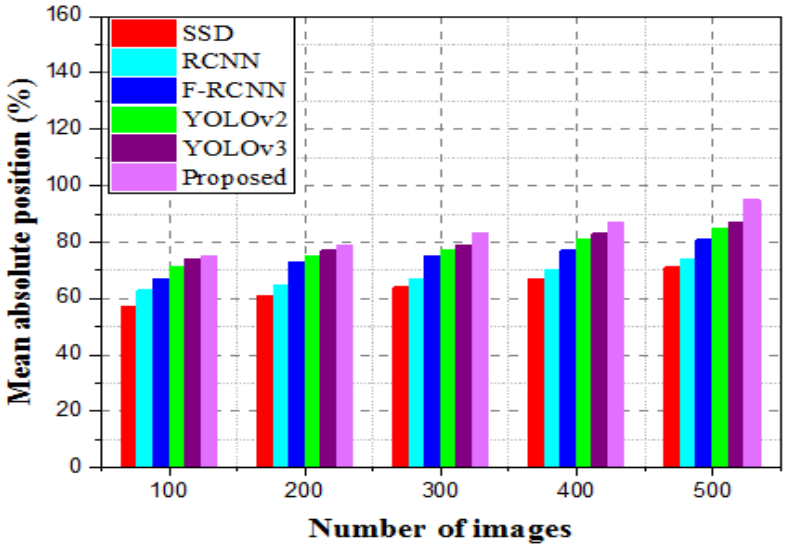

Fig 12. Analysis of Mean Absolute Position

Quantitative Results

Table 10. MAE and Computational Load Performance

Algorithms	MAE			Computation time(s)		
	Static	Dynamic	Static and dynamic	Static	Dynamic	Static and dynamic
SSD	0.37	0.39	0.35	78	76	70
RCNN	0.25	0.27	0.21	72	69	55
F-RCNN	0.18	0.20	0.16	50	47	40
YOLOv2	0.08	0.09	0.04	45	35	28
YOLOv3	0.07	0.08	0.04	30	28	22
Proposed	0.05	0.06	0.02	25	22	18

According to quantitative research, combining static and dynamic models can increase prominence detection performance. When merging static foreground networks with dynamic highlight networks, traditional approaches cannot recognize the relevance of video objects. The modelling approach is trained using static foreground data, which results in more accurate predictions than other methods. According to the above research, this work may assume that when training data drops, so performs, and vice versa. This implies that the suggested approach is data-driven. The computational load of the proposed technique and existing algorithms is compared in table 10.

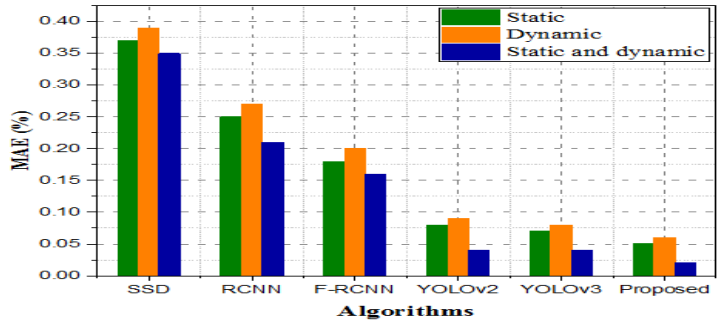

Fig 13. MAE Proposed Models Over Existing Deep Learning Models

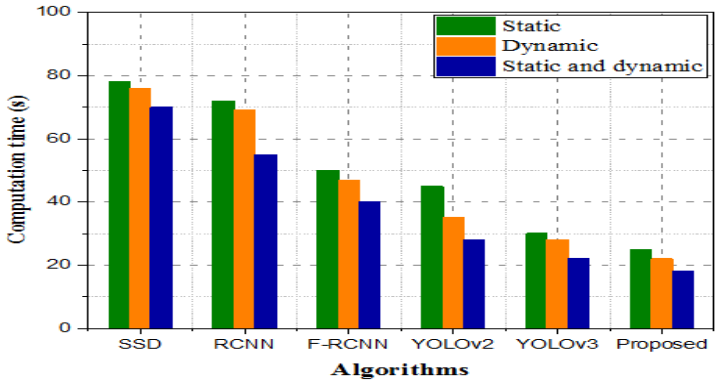

Fig 14. Computational Proposed Models Over Existing Deep Learning Models

It is clear that the suggested approach is faster than the other ways. This method has been found to reduce computation time and eliminate a significant bottleneck in the efficiency of execution. In most circumstances, motion or edge data computations impede video prominence. The outcomes are depicted in **Fig 13-14**. This encompasses both static and dynamic effects, as well as static and dynamic relationships between the suggested approaches and other conventional methods. In comparison to static or dynamic procedures and other similar methods, the suggested strategy utilizing static and dynamic procedures minimizes MAE and computing costs. Because computing time is minimized, fewer networks are offered to process incoming data.

FPGA Performance

The hardware configuration and functionality of a Xilinx Zynq-7020 FPGA at 100 MHz are directly compared to the outcomes of the algorithms. For testing, this work employs ISIM-integrated logic simulation software. After synthesis and deployment, the timing findings and latency requirements are initially analyzed to confirm that the behavior is met.

Power Consumption

Table 11 also contrasts the suggested system's energy consumption with the more advanced technique. The proposed event camera in this suggested system consumes a few watts (0.33 W), The algorithm performance alone contributes only 0.33 W of dynamic power to the device.

Table 11. Power Consumption and Latency Of Existing Object Detection Systems Versus The Proposed Method.

Algorithm	Frequency(Mhz)	Power(Watts)	Latency(ns)
Proposed	100	0.33	420
YOLOv3	100	0.37	565
YOLOv2	50	0.69	670
FRCNN	58	0.82	720
RCNN	50	0.98	760
SSD	2600	1.0	815

The current study employs the hybrid computing capabilities of Xilinx Zynq devices. However, it is limited by the high latency of frame-based systems. The Zynq module is a strong and diverse development structure; however, it can employ sleep mode and non-volatile memory, and its efficiency is significantly higher than its usefulness, with a substantially lower overall power consumption than the Smart Fusion FPGA. In other words, with correct hardware selection and construction efforts, there are numerous possibilities for the framework's low power consumption (less than 1 W). The recovery comparison utilizing the proposed approach is shown in **Fig 15**.

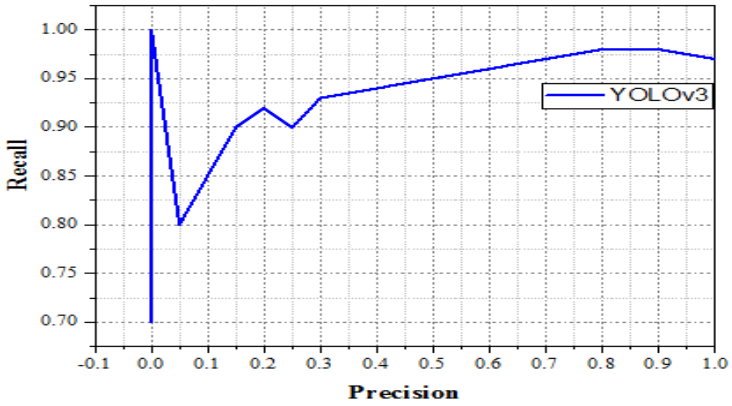

Fig 15 . Comparison Of Recall Of Proposed Algorithms

V. CONCLUSION

Traditional detection approaches cannot match the criteria for high-precision, real-time object recognition and classification in dynamic event cameras. This research creates an enhanced YOLOv3 network for thorough consideration. YOLOv3 and MobileNet were used to create a new object detection method. Initially, rather than picking fixed feature maps in the conventional YOLO v3 architecture, the technique for determining feature maps in the MobileNet is optimized based on examining the receptive fields. Experimental outcomes show that the suggested approach can identify and track foreground objects in complicated and vibrant environments with excellent precision, resilience, and efficacy. This approach also yields photos that are smooth and free of noise. Another advantage of this strategy is that it demands less computing time. The suggested method does not suffer from false object tracking even in the circumstance of varying lighting, making the system more efficient and resilient. Future work can be enhanced by experimenting with more video streams in congested areas.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Ethics Approval and Consent to Participate

The research has consent for Ethical Approval and Consent to participate.

Competing Interests

There are no competing interests.

References

[1]. L. Jiao et al., "A Survey of Deep Learning-Based Object Detection," IEEE Access, vol. 7, pp. 128837–128868, 2019, doi: 10.1109/access.2019.2939201.

[2]. R. Dixit and H. Singh, "Comparison of Detection and Classification Algorithms Using Boolean and Fuzzy Techniques," Advances in Fuzzy Systems, vol. 2012, pp. 1–10, 2012, doi: 10.1155/2012/406204.

[3]. M. N. Khan, M. Al Hasan, and S. Anwar, "Improving the Robustness of Object Detection Through a Multi-Camera–Based Fusion Algorithm Using Fuzzy Logic," Frontiers in Artificial Intelligence, vol. 4, May 2021, doi: 10.3389/frai.2021.638951.

[4]. 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, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/tpami.2016.2577031.

[5]. M. A. Rashidan, Y. M. Mustafah, A. A. Shafie, N. A. Zainuddin, N. N. A. Aziz, and A. W. Azman, "Moving Object Detection and Classification Using Neuro-Fuzzy Approach," International Journal of Multimedia and Ubiquitous Engineering, vol. 11, no. 4, pp. 253–266, Apr. 2016, doi: 10.14257/ijmue.2016.11.4.26.

[6]. S. Et. al., "Detection of Moving Vehicles on Highway using Fuzzy Logic for Smart Surveillance System," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 1S, pp. 419–431, Apr. 2021, doi: 10.17762/turcomat.v12i1s.1888.

[7]. M. Jiang, C. Deng, Z. Pan, L. Wang, and X. Sun, "Multiobject Tracking in Videos Based on LSTM and Deep Reinforcement Learning," Complexity, vol. 2018, pp. 1–12, Nov. 2018, doi: 10.1155/2018/4695890.

[8]. L. Shi, Y. Wan, X. Gao, and M. Wang, "Feature Selection for Object-Based Classification of High-Resolution Remote Sensing Images Based on the Combination of a Genetic Algorithm and Tabu Search," Computational Intelligence and Neuroscience, vol. 2018, pp. 1–13, 2018, doi: 10.1155/2018/6595792.

[9]. G. Lee, R. Mallipeddi, G.-J. Jang, and M. Lee, "A Genetic Algorithm-Based Moving Object Detection for Real-time Traffic Surveillance," IEEE Signal Processing Letters, vol. 22, no. 10, pp. 1619–1622, Oct. 2015, doi: 10.1109/lsp.2015.2417592.

[10]. N. H. Reyes and E. P. Dadios, "Dynamic Color Object Recognition Using Fuzzy Logic," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 8, no. 1, pp. 29–38, Jan. 2004, doi: 10.20965/jaciii.2004.p0029.

[11]. Z. Guo, M. Zhang, and D.-J. Lee, "Efficient Evolutionary Learning Algorithm for Real-Time Embedded Vision Applications," Electronics, vol. 8, no. 11, p. 1367, Nov. 2019, doi: 10.3390/electronics8111367.

[12]. J. H. Lee, T. Delbruck, and M. Pfeiffer, "Training Deep Spiking Neural Networks Using Backpropagation," Frontiers in Neuroscience, vol. 10, Nov. 2016, doi: 10.3389/fnins.2016.00508.

[13]. Kugele, T. Pfeil, M. Pfeiffer, and E. Chicca, "Hybrid SNN-ANN: Energy-Efficient Classification and Object Detection for Event-Based Vision," Pattern Recognition, pp. 297–312, 2021, doi: 10.1007/978-3-030-92659-5_19.

[14]. Perot, Etienne & Tournemire, Pierre & Nitti, Davide & Masci, Jonathan & Sironi, Amos. (2020). Learning to Detect Objects with a 1 Megapixel Event Camera.

[15]. B. Ramesh, A. Ussa, L. Della Vedova, H. Yang, and G. Orchard, "Low-Power Dynamic Object Detection and Classification With Freely Moving Event Cameras," Frontiers in Neuroscience, vol. 14, Feb. 2020, doi: 10.3389/fnins.2020.00135.

[16]. S. Zhang, W. Wang, H. Li, and S. Zhang, "EVtracker: An Event-Driven Spatiotemporal Method for Dynamic Object Tracking," Sensors, vol. 22, no. 16, p. 6090, Aug. 2022, doi: 10.3390/s22166090.

[17]. M. A. Perez-Cutino, A. G. Eguiluz, J. R. M. Dios, and A. Ollero, "Event-based human intrusion detection in UAS using Deep Learning," 2021 International Conference on Unmanned Aircraft Systems (ICUAS), Jun. 2021, doi: 10.1109/icuas51884.2021.9476677.

[18]. K. Srinivas, L. Singh, S. R. Chavva, B. Dappuri, S. Chandrasekaran, and S. Qamar, "Multi-modal cyber security based object detection by classification using deep learning and background suppression techniques," Computers and Electrical Engineering, vol. 103, p. 108333, Oct. 2022, doi: 10.1016/j.compeleceng.2022.108333.

[19]. S. Al-Otaibi, V. Cherappa, T. Thangarajan, R. Shanmugam, P. Ananth, and S. Arulswamy, "Hybrid K-Medoids with Energy-Efficient Sunflower Optimization Algorithm for Wireless Sensor Networks," Sustainability, vol. 15, no. 7, p. 5759, Mar. 2023, doi: 10.3390/su15075759.

[20]. B. Gökçe and G. Sonugür, "Recognition of dynamic objects from UGVs using Interconnected Neural network-based Computer Vision system," Automatika, vol. 63, no. 2, pp. 244–258, Jan. 2022, doi: 10.1080/00051144.2022.2031539.

[21]. O. A. Pakhomova and O. J. Kravets, "Efficiency analysis of dynamic object detection in computer vision system," Journal of Physics: Conference Series, vol. 1203, p. 012048, Apr. 2019, doi: 10.1088/1742-6596/1203/1/012048.

[22]. J. Duo and L. Zhao, "An Asynchronous Real-Time Corner Extraction and Tracking Algorithm for Event Camera," Sensors, vol. 21, no. 4, p. 1475, Feb. 2021, doi: 10.3390/s21041475.

[23]. J. Furmonas, J. Liobe, and V. Barzdenas, "Analytical Review of Event-Based Camera Depth Estimation Methods and Systems," Sensors, vol. 22, no. 3, p. 1201, Feb. 2022, doi: 10.3390/s22031201.

[24]. T. Ozawa, Y. Sekikawa, and H. Saito, "Accuracy and Speed Improvement of Event Camera Motion Estimation Using a Bird's-Eye View Transformation," Sensors, vol. 22, no. 3, p. 773, Jan. 2022, doi: 10.3390/s22030773.

[25]. Mohana and H. V. Ravish Aradhya, "Simulation of Object Detection Algorithms for Video Survillance Applications," 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, Aug. 2018, doi: 10.1109/i-smac.2018.8653665.

[26]. Raghunandan, Mohana, P. Raghav, and H. V. R. Aradhya, "Object Detection Algorithms for Video Surveillance Applications," 2018 International Conference on Communication and Signal Processing (ICCSP), Apr. 2018, doi: 10.1109/iccsp.2018.8524461.

[27]. K. Cai, X. Miao, W. Wang, H. Pang, Y. Liu, and J. Song, "A modified YOLOv3 model for fish detection based on MobileNetv1 as backbone," Aquacultural Engineering, vol. 91, p. 102117, Nov. 2020, doi: 10.1016/j.aquaeng.2020.102117.

[28]. Howard, Andrew & Zhu, Menglong & Chen, Bo & Kalenichenko, Dmitry & Wang, Weijun & Weyand, Tobias & Andreetto, Marco & Adam, Hartwig. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[29]. X. Li, M. Tian, S. Kong, L. Wu, and J. Yu, "A modified YOLOv3 detection method for vision-based water surface garbage capture robot," International Journal of Advanced Robotic Systems, vol. 17, no. 3, p. 172988142093271, May 2020, doi: 10.1177/1729881420932715.

[30]. D. Cao, Z. Chen, and L. Gao, "An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks," Human-centric Computing and Information Sciences, vol. 10, no. 1, Apr. 2020, doi: 10.1186/s13673-020-00219-9.

[31]. P. Dendorfer et al., "MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking," International Journal of Computer Vision, vol. 129, no. 4, pp. 845–881, Dec. 2020, doi: 10.1007/s11263-020-01393-0.