INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Signal and Imaging Systems Engineering

ISSN online: 1748-0701 - ISSN print: 1748-0698
https://www.inderscience.com/ijsise

Performance analysis of object detection and tracking methodology for video synopsis

Swati Jagtap, Nilkanth Chopade

DOI: 10.1504/IJSISE.2024.10060320

Article History:

Received:	29 May 2023
Last revised:	21 July 2023
Accepted:	09 October 2023
Published online:	15 July 2024

Performance analysis of object detection and tracking methodology for video synopsis

Swati Jagtap* and Nilkanth Chopade

Department of Electronics and Telecommunication,
Pimpri Chinchwad College of Engineering,
411044, India
Email: swatialjagtap@gmail.com
Email: nilkanth.chopade@pccoepune.org
*Corresponding author

Abstract: The enormous amount of data produced by 24/7 surveillance cameras is challenging for retrieval and browsing of video. The challenges can be overcome by reducing the video size through video condensation methods without affecting the information. Video synopsis is a condensation technique where the long video is represented in shorter form by reducing the spatial and temporal redundancy based upon the occurrence of activity that eases the video browsing and retrieval. The detection and tracking an object in a surveillance camera are essential steps in video synopsis. The proposed research compares different detection and tracking algorithms used as a first stage for video synopsis. The condensation ratio get affected due to improper detection and tracking algorithm selection. Based on evaluating both quantitative and qualitative parameters, the You Only Look Once version 4 (YOLOV4) network outperforms the Gaussian mixture model (GMM) and SSDMobileNet in detecting multiple objects within video surveillance datasets. This research will be helpful to the researcher in identifying the correct pre-processing steps in the domain of video synopsis. In future research, incorporating an auto-learning anchor model could significantly enhance accuracy.

Keywords: video synopsis; object detection; object tracking; YOLOV4; You Only Look Once version 4; GMM; Gaussian mixture model; video condensation.

Reference to this paper should be made as follows: Jagtap, S. and Chopade, N. (2024) ‘Performance analysis of object detection and tracking methodology for video synopsis’, *Int. J. Signal and Imaging Systems Engineering*, Vol. 13, No. 1, pp.10–18.

Biographical notes: Swati Jagtap received a BE degree in Electronics and Telecommunication from Shivaji University, India, in 2005 and an ME degree in Digital Systems from Pune University, India, in 2011. She is pursuing her PhD in Signal Processing from Savitribai Phule University, India. She is an Assistant Professor at the Electronic and Telecommunication Department, Pimpri Chinchwad College of Engineering, Pune, India. Her research interests include video processing, machine learning, and deep learning.

Nilkanth Chopade completed his PhD in Engineering in the year 2009. Currently, he is working as Deputy Director and Professor at the Electronic and Telecommunication Department, Pimpri Chinchwad College of Engineering, Pune, India. His research areas include signal transforms, signal processing, antenna arrays for mobile systems, and smart antennas. He has published more than 51 papers in Scopus-indexed journals and conferences. He has received over 20 lakhs of research Grants under different government schemes.

1 Introduction

As the demand for video surveillance continues to surge across various fields, it is crucial to tackle storage, monitoring, security, and browsing challenges (Elharrouss et al., 2021). The solution to overcome the difficulties is the condensation method, where the video is compressed in spatial and temporal to represent it in a shorter period. Video synopsis is a condensation method based on rearrangements of activities to show them in time-elapsed records (Jagtap and Chopade, 2021). Analysing and rearranging the sequence of activities within a video makes it possible to compress footage hours into a shorter form for storage. In video synopsis, Object detection and tracking are the first stages, and the compression ratio depends on the accuracy of its output.

Detecting any object in an image or video frame through machine vision has attracted many. Object detection has become the backbone for implementing many real-time applications such as face detection, autonomous vehicle, security, surveillance, people detection, pose detection, healthcare, robotic vision, human activity detection, and many more (Luwe et al., 2023). Object detection deals with the classification, localisation, and identification of the object in an image or video. The classification step gives the class of the objects in an image. The localisation helps to identify the position or the coordinates where the object is present. The detection step represents the bounding box bx, by, bw, bh around the object with their class. In the context of object detection, 'bx' refers to the x-coordinate of the bounding box, 'by' refers to the y-coordinate of the bounding box, 'bw' refers to the width of the bounding box, and 'bh' refers to the height of the bounding box.

The traditional handcrafted feature extraction algorithm gives good results for single object detection. Extracting various features specific to each object becomes essential to achieve multiple object detection, enhancing the classification time period and increasing complexity. Also, efficient image segmentation algorithms must be added to improve performance (Zaki et al., 2015). The deep learning approach has proven to be more effective for multiple object detection than the traditional approaches based on handcrafted features (Zhao et al., 2019).

The sequence of frames includes both stationary and moving objects. Moving objects like birds, cars, and people are foreground objects, while non-moving or static objects are background objects. The detection algorithm can be able to find the foreground objects. The challenging problem of object detection in photos and movies has driven the computer vision domain. Object-detection algorithms are susceptible to clutter, occlusion, and illumination and cannot reliably discover small things (Ahmed et al., 2021). New approaches use the neural network to reduce the challenges and increase accuracy.

Object tracking deals with the prediction of the target object from previous information. It checks the object's presence in videos and tracks the trajectories through consecutive frames. Object tracking is implemented after detecting the object's presence in the frames. Each object is assigned a unique I.D. and bounding box coordinates. This I.D. is tracked in each consecutive frame, and relevant information is stored. The trajectories are traced as per the categories such as vehicle, person, and bicycle in the frame. The complexity of tracking multiple objects exceeds that of tracking a single one, as each object needs to be assigned a unique I.D (Luo et al., 2017).

Additionally, the issues with occlusion, background noise, and pose changes are more complicated than the problems with tracking a single object. The deep learning approaches for multiple object tracking improve performance more than traditional approaches. The deep learning algorithm is proven to be more effective for object tracking related to tracking prediction and data association (Chen et al., 2019).

The paper is organised as follows: Section 2 represents our research's related work. Section 3 describes the proposed methodology. Section 4 explains the results and experiments obtained. Finally, Section 5 concludes the paper.

2 Related work

Object detection and tracking is the vital step of video synopsis, which may affect the condensation ratio and add artefacts in the output video. This section surveys different object detection and tracking algorithms used in video synopsis.

Various techniques have been proposed for object detection, including pixel difference with background cut, optical flow, frame difference, temporal median, Gaussian mixture models (GMMs), and min-cut and adaptive background modelling. Furthermore, several tracking methods, including 3D graph-cut, Euclidian distance, Kalman filter, and particle filter, are presented as multiple object tracking for tracking these detected objects across the video frames to produce the object tubes. The detection and tracking techniques determine the performance of the video synopsis because false detection and tracking have a detrimental impact on accuracy.

In the paper (Pritch et al., 2007), smooth foreground object segmentation is achieved by retrieving the moving object using background removal and min-cut. Because they are not associated with motion boundaries, image gradients that concur with background gradients are reduced. The background modelling is represented through mathematical modelling to minimise Gibb's energy.

In another research (Rodriguez, 2010), optical flow is computed for the entire sequence, and its corresponding vector is represented in the Clifford Fourier domain. The flow estimation method is used to identify the dynamic regions of the video sequence. This region is considered a potential candidate, and its spatiotemporal locations are included in the final video summary.

A novel online video synopsis method is proposed (Huang et al., 2012) to solve the problem of video browsing in extensive storage data by rearranging position and foreground objects in consecutive frames in real-time chronological order. The method can be directly applied to streaming surveillance videos to obtain synopsis video. A descriptor-based appearance model with the motion model is used to extract and track the foreground object from the video. The synopsis table is proposed to store the positions of the foreground object in online mode. The enhanced Contrast Context Histogram determines whether the object is being tracked or corresponds to a new item, facilitating the reordering of positions in the resulting video. This modified histogram demonstrates robustness against both photometric and geometric transformations.

The paper (Yao et al., 2014) implemented a frame-differencing background modelling algorithm to extract the background frame. The fixed interval of frames is subtracted from adjacent frames to get frame-difference images with the object's position. The average of all images is used to get the final background frame. Kalman filter is used to predict the location of moving objects, and the matching model is generated using segmented objects. The original image is divided into four search areas based on the motion direction, distance from the centre, and matching priority.

As occlusion affects the output of object detection, the paper (Lee and Ko, 2005) proposes occlusion activity detection and prediction by the Kalman filter. In this, the labelled blob of current occlusion between the objects is predicted using a Kalman filter, and occlusion status is updated to decide the frequency of occlusion.

Online conventional video synopsis system does not require pre-processing to build tubes and background graphics. It segments moving objects using scale-invariant local ternary pattern (SILTP) feature-based background subtraction (Zhu et al., 2015).

The visual background extractor (ViBe) algorithm is used for background creation and tube extraction (Barnich and Droogenbroeck et al., 2011). ViBe is an effective and all-encompassing sample-based background subtraction technique that determines the background by comparing its value to neighbour values in a sphere. The linear neighbour object prediction algorithm is used to extract tubes, and these tubes help for extracting the items (He et al., 2017)

In the paper (He et al., 2018), moving objects are identified using the GMM, and linear prediction-based tracking is used to derive object trajectories. The rearranged events are inserted into the background image, which can be retrieved using temporal median or moving average in temporal space to create the summarised video once the spatiotemporal information of each object is collected.

In Ra and Kim (2018), for the video synopsis optimisation method, the GMM (Zivkovic, 2004) is used to distinguish the object's foreground from the backdrop during the background modelling stage. The Hungarian method (Kuhn, 1955) is used to construct the object tubes, which associate foregrounds that belong to the same object. The created object tubes are saved in a queue and stored for later use. The tube rearrangement algorithm sets the first object tube labels in the queue. The queue's contents are cleared when its size exceeds a certain threshold.

In the paper (Nie et al., 2020), the background estimation of surveillance video is computed using a median temporal filter for every minute of the video to accommodate the illumination changes. The moving object is extracted by subtracting the original frame from the corresponding background frame, followed by morphological operations. The GMM is used for more complex illumination conditions for background subtraction.

In the research (Ahmed, 2020), The GMM is used to estimate the density of these components. A bounding box represents the detected object, and the Hungarian optimisation method is used to match each detection with an object. The lost objects in consecutive frames are identified using Kernelised Correlation Filter. Kalman filter is used to reduce the occlusion. The classification of moving objects is represented into three categories, i.e., bike, car, and pedestrians, using three layers of Convolution Neural Network.

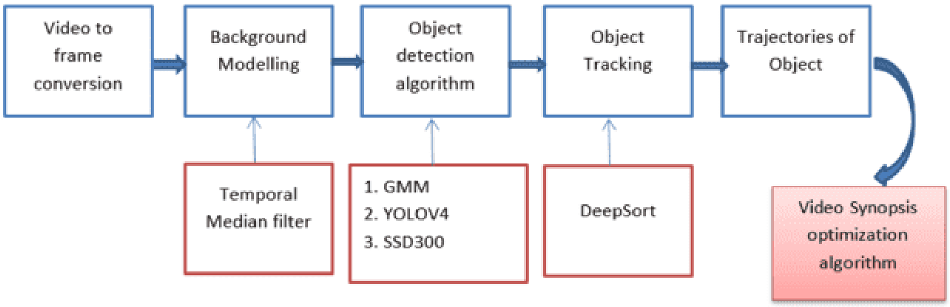
The paper (Moussa et al., 2021) uses deep learning approaches such as YOLOv3 to detect objects. The object tracking uses a simple online real-time tracking algorithm with deep association metric DeepSort. The Hungarian algorithm is used to associate the object track with detection. The prediction of the next state of the object is made by Kalman filter based on the bounding box. For visualisation and query processing, object tracking creates 2D trajectories that are merged with object properties (Namitha et al., 2022). Poisson image editing (Perez et al., 2003) stitches the rearranged tubes to the foreground.

The traditional methods (Girshick et al., 2014) of object detection that use feature vector creation and extraction are limited to detecting a single category of objects, while deep learning algorithms are the breakthrough into modern algorithms with multiple detection.

3 Methodology

The suggested investigation plays a crucial role as an initial step in video synopsis, making it an essential contribution to the field. Properly selecting detection and tracking algorithms helps reduce the visual artefacts and achieves a good condensation ratio. A comparative analysis of five different algorithms is proposed, including background subtraction, Optical flow, GMM, You Only Look Once version 4 (YOLOV4), and SSDMobileNet. For object tracking, DeepSort is used.. The flow of research is as follows.

Figure 1 Methodology (see online version for colours)



In video synopsis, the condensation ratio depends on the output of object detection, which is the first and most crucial step, followed by tracking and optimisation. The foreground mask of the static video needs to be extracted precisely and quickly for real-time applications.

For surveillance applications, the CCTV footage has static backgrounds. After converting video into frames, the static background is extracted using a temporal median filter applied to the frames sequence. Due to the surveillance camera's round-the-clock operation, it is common to encounter numerous frames where no activity is detected. To reduce the computational complexity, we remove frames that do not contain any detected objects.

The moving object in the frames of static cameras can be detected using many approaches. Background subtraction is a traditional approach that relies on finding the difference between the foreground mask and the current frame. But the method can only detect moving objects without having any information about the object type and other details. Optical flow is another approach to detecting objects in motion between consecutive frames. Optical flow determines a point's velocity within a frame and predicts potential locations for points in the following successive frame sequence.

Gaussian mixture model can be used to model the background of a scene and distinguish foreground objects from the background. The basic idea behind using GMM for object detection is first to model the background of the scene using a mixture of Gaussian distributions. This is done by analysing the colour and intensity values of the pixels in the scene and fitting a set of Gaussian distributions to them. Once the background model has been constructed, it can detect objects in the scene. Any pixels that do not match the background model are considered to be part of a foreground object. While Gaussian Misture Model-based object detection can be effective in specific scenarios, it does have some limitations. For example, it can struggle with scenes where the background is not stationary, such as a waving flag or a moving tree. It can also struggle with scenes where there is significant lighting variation or shadows. Other object detection techniques, like deep learning-based approaches, may be more effective.

YOLOV4 is a state-of-the-art deep learning model for object detection that uses a convolutional neural network (CNN) to perform object detection and classification. The model outputs a set of bounding boxes representing the locations of the detected objects in the image, their corresponding class labels, and confidence scores. YOLOV4 offers several advantages over other object detection models, including high accuracy, speed, and robustness to variations in lighting, occlusion, and object size (Bochkovskiy et al., 2020).

SSDMobileNet (Single Shot Detector with MobileNet architecture) is a popular object detection algorithm that detects and classifies objects in images. It achieves high accuracy by combining convolutional and pooling layers to extract features from the input image. It also uses a multi-scale detection strategy that performs object detection at multiple scales in the image. It achieves high speed using a single-shot detection approach that detects objects in a single pass through the network. Finally, it achieves robustness to object size and aspect ratio variations by using a set of predefined anchor boxes optimised for different object sizes and aspect ratios (Liu et al., 2016).

DeepSort (Deep Learning with simple online real-time tracking) is a deep learning-based object tracking algorithm that combines a deep appearance descriptor with a Kalman filter-based state estimation algorithm. DeepSORT aims to track objects in a video sequence by associating detections of the same object over time while providing a unique identification number to each object. The combination of deep appearance descriptors and Kalman filtering enables Deep SORT to handle occlusions, appearance changes, and missing detections. Overall, DeepSort provides a robust and efficient object-tracking framework in real-world scenarios. (Wojke et al., 2017)

4 Results and discussion

The experiment setups are conducted to perform a comparative analysis of various object detection algorithms. The output of object detection and tracking affects the performance of the video synopsis. One of the evaluation parameters of video synopsis is the condensation ratio. It is the ratio of frames in a synopsis video to the number of frames in the original video. Ideally, the condensation radio should be as small as possible. The condensation ratio increases with the improper selection of detection and tracking algorithms. Also, the visual artefacts in the output video synopsis give an unrealistic view. The performance is evaluated on the system specification of Intel core i5, RAM 8 GB, Nvidia GEFORCE GTX 1050. The dataset used is eight videos of the VIRAT (Oh et al., 2011) dataset and some of the videos from MOT17 (Sun et al., 2019). These datasets are the benchmark in computer vision, with surveillance video covering diversity in terms of resolution, background clutter, and many more parameters.

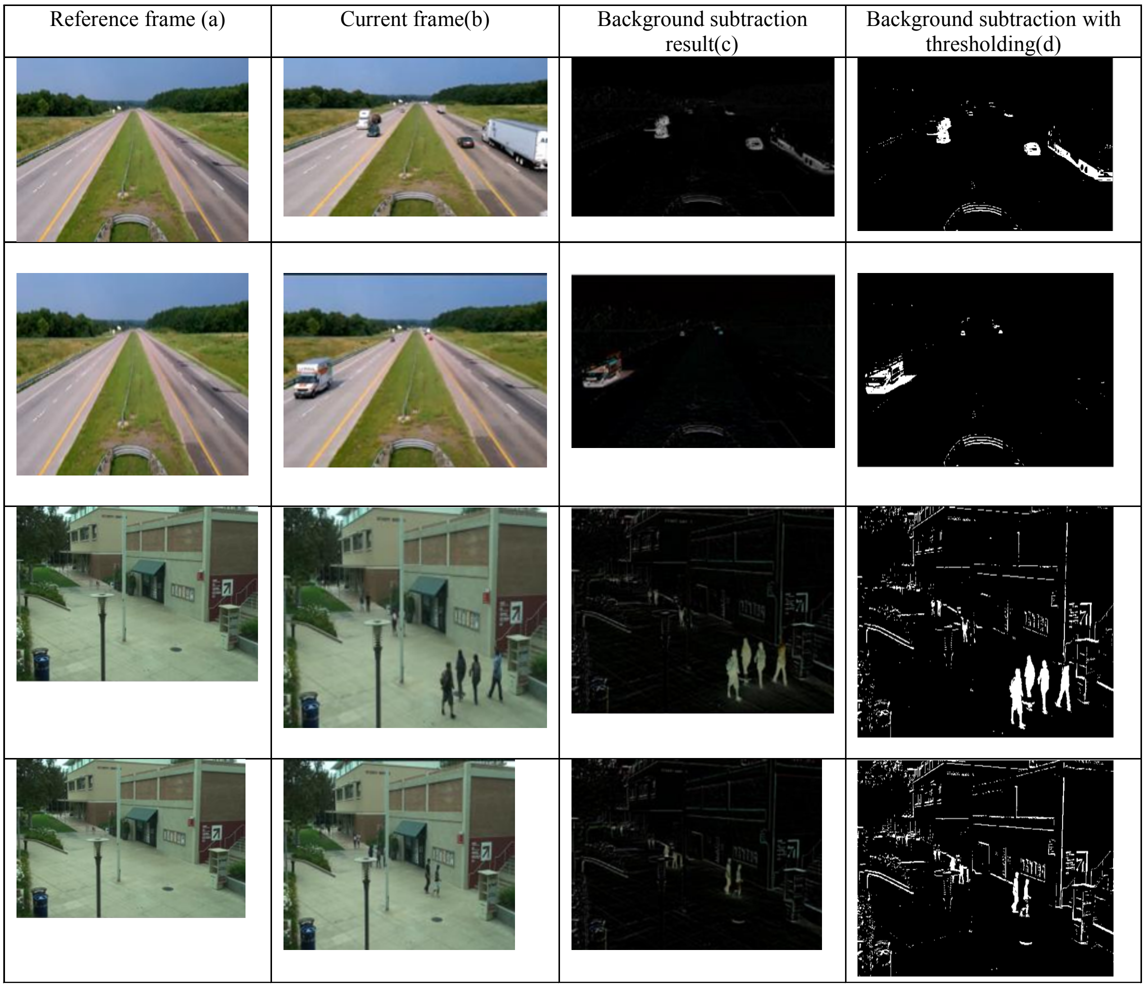
The performance of an object detection algorithm is assessed using several parameters such as Precision, Recall, F1 score, Average Precision, Intersection over union, and Mean Average Precision. These evaluation parameters help assess object detection algorithm's performance and identify improvement areas. The number of images that are processed per second is called Frames Per Second. It evaluates the speed of the detection model.

Background subtraction helps to identify the presence of moving objects in the surveillance static camera. The background image is used as a reference image and is extracted by taking a temporal average of all the frames. This method finds the difference between a reference image and a current image, and based on the result of the subtraction, it gives information about the current object in the frame.

Figure 2 gives the result of background subtraction and detection of the presence of a moving object. The main challenge in this method is that the filling of the object is not correctly identified, leading to changes in the object's shape. Also, it is challenging to identify the type of the object.

Optical flow is another method used to detect the object in motion by calculating the optical flow of the image, which gives the velocity of the image pixel and provides the estimation of where the point may be in the sequence of the next frame. Each vector is represented by direction and magnitude information, which helps track the object.

Figure 2 (a) Reference frame; (b) current frame; (c) result of background subtraction and (d) result of background subtraction with thresholding (see online version for colours)


Reference frame (a)
Current frame(b)
Background subtraction result(c)
Background subtraction with thresholding(d)

The presence of moving objects is easily traced by optical flow, as shown in the figure. It also gives the direction with the velocity of motion vectors. The computation complexity increases with the number of objects and their sensitivity to noise.

Another background subtraction algorithm used is the GMM to detect objects. It's the mixture of different Gaussian curves based on pixel intensity and plotted through a histogram. It is an unsupervised clustering algorithm. Each Gaussian has its mean, standard deviation, and width. The background pixels are represented as no change in intensity value, while the foreground pixels are denoted as changes in intensity in the histogram curve.

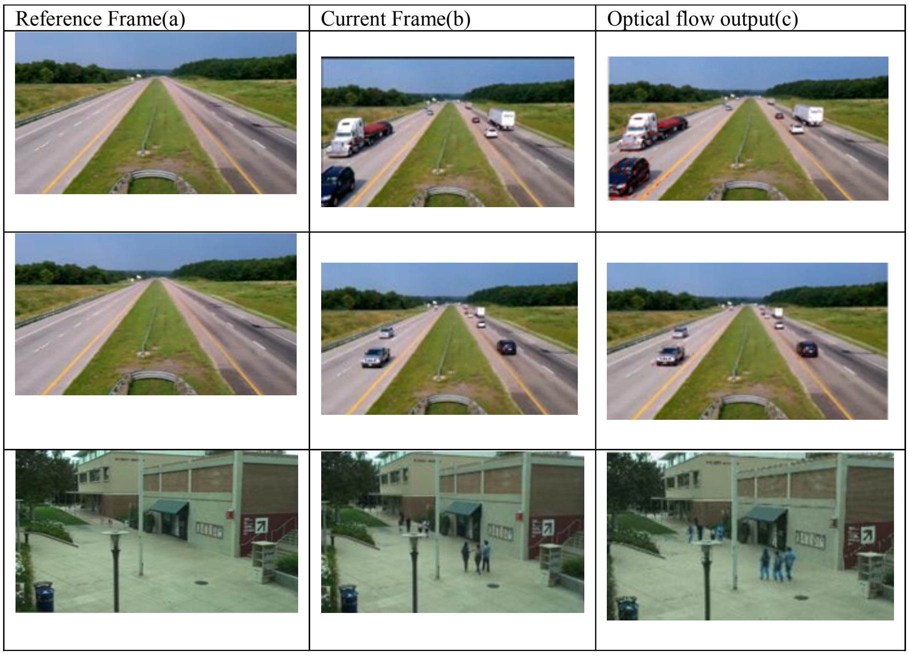
Figure 4 gives the result of the GMM applied to the vehicle dataset, which compares the ground truth labelled frame with the detected frame. In Figure 4(b), six cars and two trucks are there, but in the detected frame, two vehicles and two trucks are detected.

Table 1 gives the Precision, Recall, and F1-score of GMM calculated on different frames of vehicle video. From the observation, GMM gives an average Precision of 80.08%, an average Recall of 65.55%, and an average F1-score of 70.81%.

Table 1 Result of the Gaussian mixture model

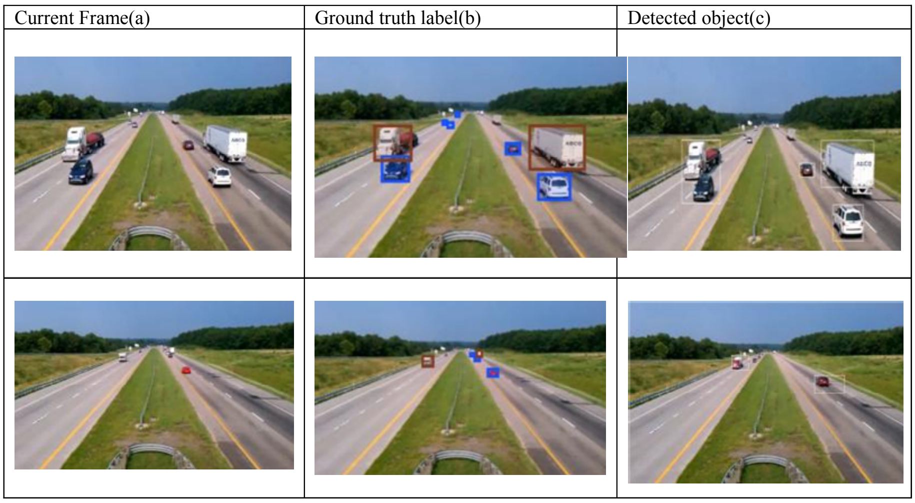
S. no.	*Video title*	*Resolution*	*Precision (%)*	*Recall (%)*	*F1-score (%)*
1	Highway Vehicle	800×480	90.90	66	76.92
2	Persons	1280×720	91.66	74.57	81.48
3	People with shadow	1920×1040	60.86	38.33	44.87
4	traffic	640×360	76.92	83.33	80
	Average		80.08	65.5575	70.8175

Figure 3 (a) Reference frame; (b) current frame and (c) snapshot of result of optical flow (see online version for colours)



Reference Frame(a) | Current Frame(b) | Optical flow output(c)

Figure 4 (a) Current frame; (b) ground truth and (c) output of Gaussian mixture model (see online version for colours)



Current Frame(a) | Ground truth label(b) | Detected object(c)

The YOLOV4 with architecture VGG 16 is implemented with the same dataset and hardware configuration using Python 3.4.

Figure 5 gives the results of YOLOV4 applied to the vehicle dataset. The frame shows the ground truth result and detected object with a confidence score.

Table 2 gives the result of YOLO V4 applied on different videos with different resolutions. The average Precision calculated is 88.62%, the average Recall is 83.36%, and the average F1 score is 85.22%.

Figure 6 details the object detection using SSDMobileNet applied on a different dataset. The three frames compare the current, Ground Truth, and Detected object frames.

Table 2 Result of YOLOV4

S. no.	*Video title*	*Resolution*	*Precision*	*Recall*	*F1-score*
1	Highway Vehicle	800×480	83.33	76.92	80.00
2	Persons	1280×720	95.23	86.96	90.90
3	People with shadow	1920×1040	94.11	69.56	80.00
4	Traffic	640×360	81.81	100	90.01
	Average		88.62	83.36	85.22

Figure 5 (a) Current frame; (b) ground truth frame and (c) output of YOLOV4 (see online version for colours)

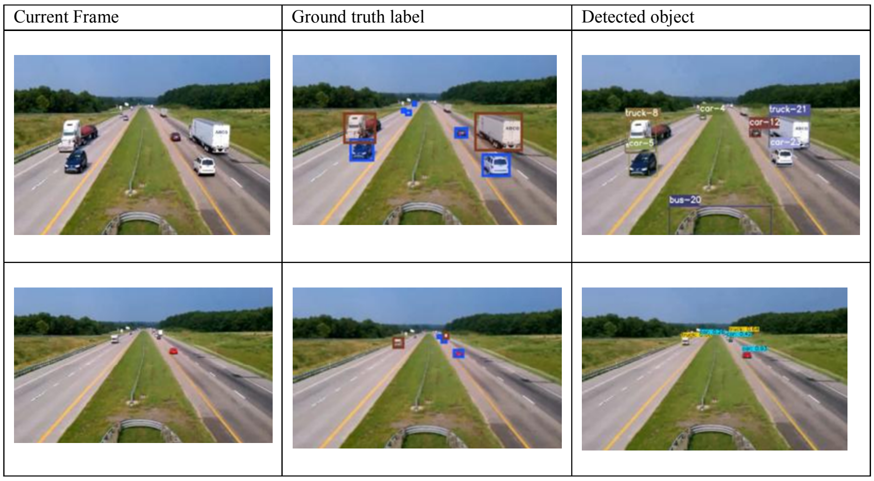


Figure 6 (a) Current frame; (b) ground truth frame and (c) output of SSDMobileNet (see online version for colours)

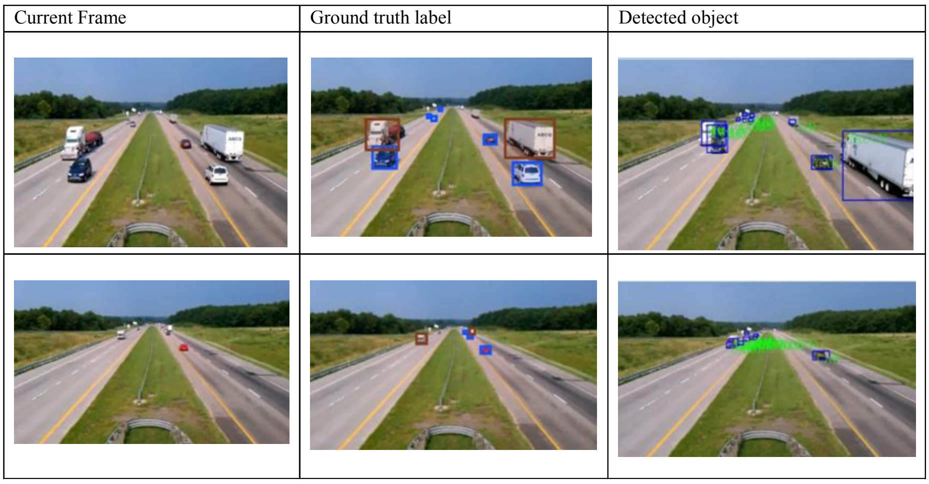


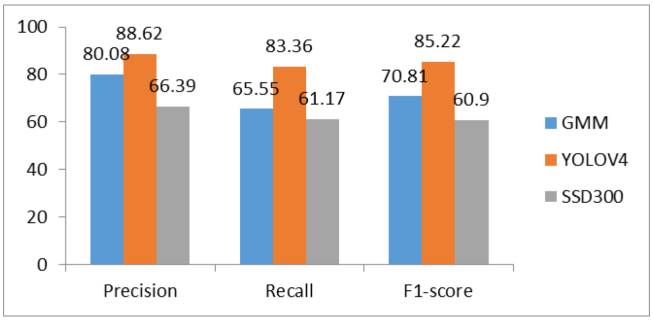
Table 3 gives the result of SSD MobileNet applied on different videos with different resolutions. The average Precision calculated is 66.39%, the average Recall is 61.17%, and the average F1 score is 60.90%.

Table 3 Precision, recall, and F1 score of SSDMobileNet

S. no.	*Video title*	*Resolution*	*Precision*	*Recall*	*F1-score*
1	Highway vehicle	800×480	53.82	77.70	63.59
2	Persons	1280×720	90.60	51.78	65.90
3	People with shadow	1920×1040	42.85	25.00	31.57
4	Traffic	640×360	78.32	90.23	82.54
	Average		66.39	61.17	60.90

Figure 7 shows the comparison results of the GMM, YOLOV4, and SSDMobileNet.

Figure 7 Comparison of an object detection algorithm for video synopsis (see online version for colours)



The observation of the experiment can be concluded as

1 Gaussian mixture model (GMM) is a statistical model used for clustering and density estimation, and it is not explicitly designed for object detection tasks. Therefore, it is unsuitable for object detection tasks compared to modern object detection algorithms such as YOLOv4 and SSDMobileNet 300.

2 YOLOv4 and SSDMobileNet 300 are state-of-the-art object detection models with different characteristics. YOLOv4 is known for its speed and accuracy in detecting objects, while SSD 300 is also fast but tends to perform better on smaller objects.

3 From the experiment, YOLOV4 gives a good result compared to the other two algorithms. The Precision, Recall, and F1 score of YOLOV4 is the highest compared to the other two. The high-computing GPU can be used to enhance the speed of YOLOV4.

5 Conclusion

Object detection and tracking algorithms are used to localise and track the abnormalities to cope with the security concerns with the exponential growth of surveillance video. The proposed research compares various object detection algorithms to identify the best-suited algorithm for video synopsis. The output of synopsis depends upon the output of detection and tracking. The comparison result shows that the pertained network, such as YOLOV4 gives around 85.44% of the F1 score compared to the GMM and SSDMobileNet. The YOLOV4 object detection model is combined with the tracking algorithm DeepSort.

The proposed research concluded that YOLOV4 gives better results than other models. As YOLOV4 is a single-stage detector, the accuracy for detecting small objects is significantly less. Also, another concern is the speed of the detector. To accelerate the inference time of the model, the main approach is to opt for a smaller model, like YOLOv4-tiny. Additionally, further improvements in inference time are achieved by selecting appropriate hardware, such as a GPU. In future research, the auto-learning anchor model can be added to improve the accuracy.

References

Ahmed, M., Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D. and Afzal, M.Z. (2021) 'Survey and performance analysis of deep learning based object detection in challenging environments', *Sensors*, Vol. 21, p.5116.

Ahmed, S.A. (2020) 'Query-based video synopsis for intelligent traffic monitoring applications', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21, pp.3457–3468.

Barnich, O. and Droogenbroeck, M. (2011) 'ViBe: a universal background subtraction algorithm for video sequences', *IEEE Transactions on Image Processing*, Vol. 20, pp.1709–1724, 10.1109/TIP.2010.2101613.

Bochkovskiy, A., Wang, H-Y. and Liao, M. (2020) *YOLOv 4: Optimal Speed and Accuracy of Object Detection*, ArXiv, https://arxiv.org/pdf/2004.10934

Chen, S., Xu, Y., Zhou, X. and Li, F. (2019) 'Deep learning for multiple object tracking: a survey', *IET Computer Vision*, Vol. 13, No. 4, pp.355–368.

Elharrouss, O., Almaadeed, N. and Al-Maadeed, S. (2021) 'A review of video surveillance systems', *Journal of Visual Communication and Image Representation*, Vol. 77, pp.103–116.

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) 'Rich feature hierarchies for accurate object detection and semantic segmentation', *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, pp.580–587.

He, Y., Han, J., Sang, N., Qu, Z. and Gao, C. (2018) 'Chronological video synopsis via events rearrangement optimization', *Chinese Journal of Electronics*, Vol. 27, pp.399–404.

He, Y., Qu, Z., Gao, C. and Sang, N. (2017) 'Fast online video synopsis based on potential collision graph', *IEEE Signal Processing Letters*, Vol. 24, pp.22–26.

Huang, C-R., Chen, H-C. and Chung, P-C. (2012) 'Online surveillance video synopsis', *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp.1843–1846, DOI: 10.1109/ISCAS.2012.6271627.

Jagtap, S. and Chopade, N.B. (2021) 'A comprehensive investigation about video synopsis methodology and research challenges', *Inventive Computation and Information Technologies*, pp.911–923, DOI:-10.1007/978-981-33-4305-4_66.

Kuhn, H.W. (1955) 'The hungarian method for the assignment problem', *Naval Res. Logistic Quart.*, Vol. 2, pp.83–97.

Lee, H. and Ko, H. (2005) 'Occlusion activity detection algorithm using kalman filter for detecting occluded multiple objects', in Sunderam, V.S., van Albada, G.D., Sloot, P.M.A. and Dongarra, J.J. (Eds.): *Computational Science – ICCS 2005. ICCS 2005*, Lecture Notes in Computer Science, Vol. 3514, Springer, Berlin, Heidelberg, https://doi.org/10.1007/11428831_18

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C-Y. and Berg, A.C. (2016) 'SSD: single shot multiBox detector', *Lecture Notes in Computer Science*, Vol. 9905, pp.21–37.

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Zhao, X. and Kim, T-K. (2017) 'Multiple object tracking: a literature review', *Artificial Intelligence*, p.293, 10.1016/j.artint.2020.103448.

Luwe, Y.J., Lee, C.P. and Lim, K.M. (2023) 'Wearable sensor-based human activity recognition with ensemble learning: a comparison stud', *International Journal of Electrical and Computer Engineering,* Vol. 13, No. 2088, pp.4029–4040.

Moussa, M. and Shoitan, R. (2021) 'Object-based video synopsis approach using particle swarm optimization', *Signal Image and Video Processing*, p.15, 10.1007/s11760-020-01794-1.

Namitha, K., Narayanan, A. and Madathilkulangara, G. (2022) 'Interactive visualization-based surveillance video synopsis', *Applied Intelligence*, p.52, 10.1007/s10489-021-02636-4.

Nie, Y., Li, Z., Zhang, Z., Zhang, Q., Ma, T. and Sun, H. (2020) 'Collision-free video synopsis incorporating object speed and size changes', *IEEE Transactions on Image Processing*, Vol. 29, pp.1465–1478.

Oh, S., Hoogs, A., Perera, A.G.A., Cuntoor, N., Chen, C-C., Lee, J, Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H. Ramanan, D., Yuen, J. and Desai, M. (2011) 'A large-scale benchmark dataset for event recognition in surveillance video', *Proceedings of IEEE Comptuer Vision and Pattern Recognition* (*CVPR*), pp.3153–3160, 10.1109/ CVPR.2011.5995586.

Perez, P., Gangnet, M. and Blake, A. (2003) 'Poisson image editing', *ACM Transactions on Graphics*, Vol. 22, p.313.

Pritch, Y., Rav-acha, A., Gutman, A. and Peleg, S. (2007) 'Webcam synopsis: peeking around the world', *IEEE International Conference on Computer Vision*, pp.1–8, doi-10.1109/ICCV.2007.4408934.

Ra, M. and Kim, W-Y. (2018) 'Parallelized tube rearrangement algorithm for online video synopsis', *IEEE Signal Processing Letters*, Vol. 25, pp.1186–1190.

Rodriguez, M. (2010) 'CRAM: compact representation of actions in movies', *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp.3328–3335, doi: 10.1109/CVPR.2010.5540030.

Sun, S., Akhtar, N., Song, H., Mian, A. and Shah, M. (2019) 'Deep affinity network for multiple object tracking', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp.104–119.

Wojke, N., Bewley, A. and Paulus, D. (2017) 'Simple online and realtime tracking with a deep association metric', *Proc. Int. Conf. on Image Processing,* Beijing, China, pp.3645–3649.

Yao, T., Xiao, M., Ma, C., Shen, C. and Li, P. (2014) 'Object based video synopsis', *Proceedings – 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications, WARTIA 2014*, pp.1138–1141, 10.1109/ WARTIA.2014.6976479.

Zaki, M., Shaheen, S. and El-Marakeby, H. (2015) 'Multiple object detection and localisation system using automatic feature selection', *International Journal of Signal and Imaging Systems Engineering*, Vol. 8, No. 3, p.146.

Zhao, Z-Q., Zheng, P., Xu, S-T. and Wu, X. (2019) 'Object detection with deep learning: a review', *IEEE Transactions on Neural Networks and Learning Systems*, pp.1–21, 10.1109/ TNNLS.2018.2876865.

Zhu, J., Feng, S., Yi, D., Liao, S., Lei, Z. and Li, S.Z. (2015) 'High-performance video condensation system', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 25, pp.1113–1124.

Zivkovic, Z. (2004) 'mproved adaptive Gaussian mixture model for background subtraction', *Proc IEEE 17th Int. Conf. Pattern Recog*, Vol. 2, pp.28–31.