Scholars Journal Category: Technology

Published By Scholar Indexing Society
E-ISSN: 2228-837X----DOI URL: http://doi.org/10.53075/Ijmsirq/665776577677656
Journal Homepage: https://scholars.originaljournals.com/ojs/index.php/ojs/index

Implementation of Object Detection and Tracking by Using Deep Learning Based Convolutional Neural Networks

Atianashie Miracle A.
Catholic University College of Ghana, Fiapre, P.O. Box 363, Sunyani
miracleatianashie81@gmail.com
ORCID: https://orcid.org/0000-0003-2217-3298

Submitted: 12/12/2021 Revised: 28/01/2022 Accepted: 14/02/2022

Abstract: Video object detection plays a significant role in various applications, including security, remote sensing and hyperspectral. In recent years, deep learning-based algorithms have made significant advances in video object recognition. The conventional machine learning applications are resulted in poor accuracy. In this article, a unified deep learning-based convolutional neural network (DLCNN) is developed for composite multi-object recognition in videos. DLCNN analyses a composite item as a collection of background and adds part information into feature information to enhance hybrid object recognition. Correct component information may help forecast the shape and size of a feature data, which helps solve challenges caused by different forms and sizes of various objects. Finally, the DLCNN draws a bounding box to detect objects using background features. Further, the simulation results show that the proposed method's performance is improved compared to the state of art approaches.

Keywords: deep learning-based convolutional neural network, object detection
DOI: 10.53075/Ijmsirq/665776577677656

I. Introduction

In recent years, deep learning-based object recognition techniques [1] that developed from computer vision have grabbed the public's interest. Object recognition methods based on deep learning frameworks have quickly become a popular way to interpret moving frames acquired by drones [2]. Most of the works used a publicly available dataset to test our technique's detection performance and generalization capabilities on three common kinds of composite items. Meanwhile, we're working on a more difficult dataset regarding sewage treatment plants, a common type of complicated composite object [3]. It alludes to the pertinent data gathered from authorities and professionals. In a point-cloud, three-dimensional objects are often represented as 3D boxes. Although there are several differences, this representation is similar to the well-studied image-based 2D bounding-box detection [4]. Objects do not have a fixed orientation in a three-dimensional world, and box-based detectors have trouble counting all of them or fitting an axis-aligned bounding box to rotating objects [5].

Instead, this work suggests that 3D objects be represented, detected, and tracked as points in this study. CenterPoint is a framework that uses a key point detector [6]to detect the centers of objects and then regresses to additional qualities such as 3D size, 3D orientation, and velocity. Second, the bounding boxes are clustered using the k-means clustering technique. The anchoring is improved based on the clustering results. The revised video frame approaches the dataset's true bounding box [7]. It refines these estimations in a second stage by incorporating additional point characteristics on the item. 3D object tracking is reduced to greedy closest-point matching in CenterPoint. The performance of object detectors and trackers has substantially increased because to the rapid growth of deep learning networks and GPU processing capability [8]. This work includes advanced deep learning models, which have recently gained popularity. Primarily, we've given you a detailed review of a wide range of generic and customized object detection models. This work compiled a list of comparison findings to help you choose the finest detector, tracker,

Published By Scholar Indexing Society
E-ISSN: 2228-837X----DOI URL: http://doi.org/10.53075/ljmsirq/665776577677656
Journal Homepage: https://scholars.originaljournals.com/ojs/index.php/ojs/index

and combination. This work also included an inventory of old and innovative object recognition and tracking software and development trends. When gazing around, humans can recognize and track nearby items by forming a spatial-temporal memory of the objects. This paper provides a unique method for integrated 3D object recognition and tracking, which employs a dynamic object occupancy map and prior object states as spatial-temporal memory to aid item detection in subsequent frames. Rest of the article is organized as follows, section 2 deals with the existing works with problems, section 3 deals with the detailed analysis of proposed method, section 4 deals with the results and discussions, and section 5 concludes the paper with possible future scopes.

2. Literature survey:

This memory guides the detector and the ego-motion from back-end odometry to achieve more efficient object proposal creation and more accurate object state estimation. Existing Multiple-Object detection and Tracking (MODT) approaches either use the tracking-by-detection paradigm to perform object detection, feature extraction, and data association independently or combine two of the three subtasks to provide a partly end-to-end solution. In [9] authors presented Chained-Tracker, a simple online model that naturally incorporates all three subtasks into an end-to-end solution (the first as far as we know). It connects paired bounding box regression results calculated from overlapping nodes, each of which spans two adjacent frames. Object-attention (provided by a detection module) and identity-attention are used to make the paired regression more attentive (ensured by an ID verification module). MIN [10] authors presented several algorithms for MODT and tracking have been developed as a result of breakthroughs in the field of machine learning as video surveillance has recently received a lot of interest in a variety of real-time applications. A novel MODT approach is introduced in this work. The suggested method employs an effective Kalman filtering algorithm to track moving objects in video frames. In [11] authors presented the region growth model was used to translate the video clips into morphological processes based on the number of frames. After differentiating the objects, the probability-based grasshopper method was used to apply Kalman filtering for parameter optimization. A similarity measure was used to monitor the selected items in each frame using the optimal settings. Finally, the suggested MODT framework was put into action, and the outcomes were evaluated.

In [12] authors presented the object detection method, which eliminates superfluous frames with minimal information after syncing numerous movies. Finally, the tracking information is used to identify the discovered product's purchase activity. An end-to-end recognition framework was created using these procedures. In addition, the suggested object detection network performs similarly to state-of-the-art approaches. In [13], authors discovered tremendous advancements in object identification, localization, and tracking in key computer vision applications. However, no approaches for detecting, localizing, and tracking objects in road contexts that consider real-time restrictions are currently available. Then, using an Intel RealSense frame, we demonstrate our depth estimation method. Finally, as the third and final phase in our process, we offer our SORT-based 3D object tracking technique. In [14] authors conducted many trials in a controlled indoor environment to validate all of the advancements. Our own dataset, which includes doors and door handles, is used to test detection, distance estimates, and object tracking. Object detection is a critical perceptual job in autonomous driving and advanced driver assistance systems. Although the visible frame is commonly employed for perception, it is restricted by lighting and environmental fluctuations.

In [15] authors presented a deep learning framework for successful sensor fusion of the visible frame with additional sensors for robust vision-based perception. A feature-level sensor fusion approach based on skip connection is presented for sensor fusion of the visible frame with the millimeter-wave video and the thermal frame. The RV-Net and the TV-Net are the names of the two networks. There are two input branches and one output branch in these networks. The individual sensor features extraction branches exist in the input branches, which are subsequently fused in the output perception branch utilising skip connections. The RVNet and the TVNet simultaneously execute sensor-specific feature extraction, feature-level fusion, and object detection within an end-to-end architecture. On the other hand, Existing tracking-by-detection algorithms frequently seek objects in every new video frame from start without fully utilizing memory from earlier detection results.

Published By Scholar Indexing Society
E-ISSN: 2228-837X----DOI URL: http://doi.org/10.53075/ljmsirq/665776577677656
Journal Homepage: https://scholars.originaljournals.com/ojs/index.php/ojs/index

3. Proposed Method

This section gives a detailed analysis of video object detection and tracking using deep learning models.

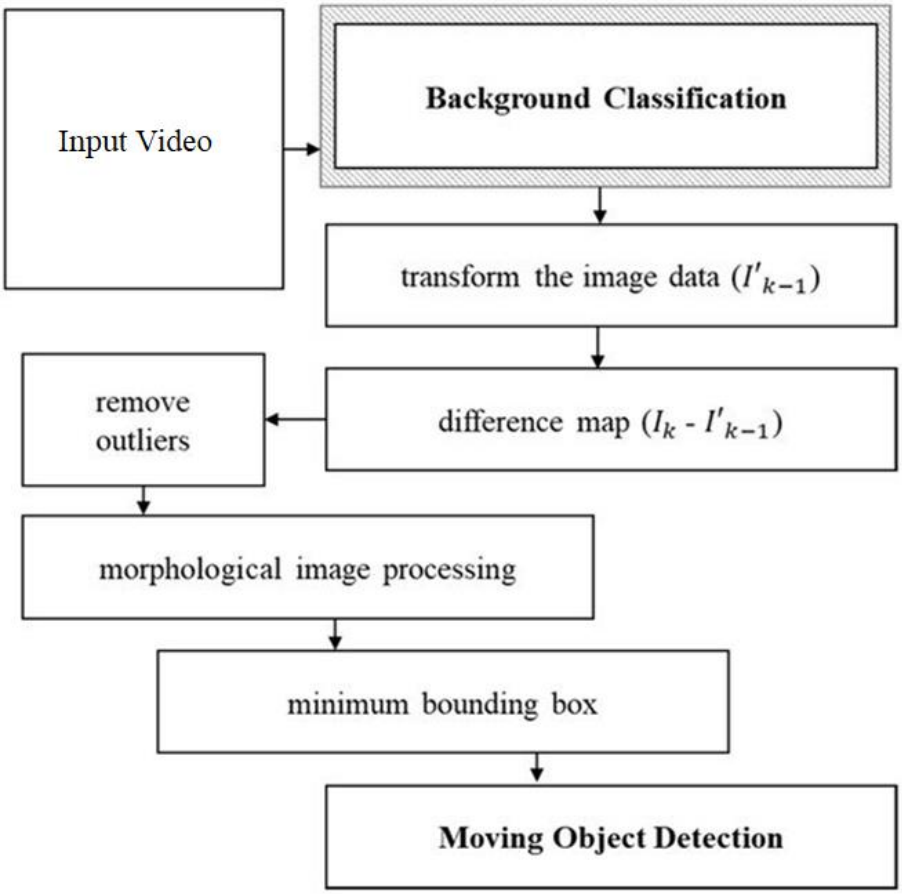


Figure 1: Proposed moving object detection system

Figure 1 is block diagram of proposed method, which is used to recognize a moving object in frame sequences of video. A DLCNN classifier is utilized to get the interesting spots in the image, which is commonly employed in image registration. The feature point correspondences in two successive frames are then computed. The related feature points may simply be utilized to generate the transformation matrix of two successive frames in well-known ways. However, as previously stated, moving object feature points might be included in the corresponding feature points, which can negatively impact the correctness of the transformation matrix computation.

It might be difficult to tell whether or not a feature point belongs to a moving item. A prior study employed the distance between the epipolar line and the feature point to generate the categorization criterion for feature points. In stereo vision systems, epipolar geometry is widely employed to determine the correspondence of the two frames from each frame. However, epipolar geometry is applied in the moving mono frame's sequential frames in the suggested technique. The basic matrix with the accompanying feature points of consecutive frames is typically calculated using a normalized eight-point technique.

Published By Scholar Indexing Society
E-ISSN: 2228-837X----DOI URL: http://doi.org/10.53075/ljmsirq/665776577677656
Journal Homepage: https://scholars.originaljournals.com/ojs/index.php/ojs/index

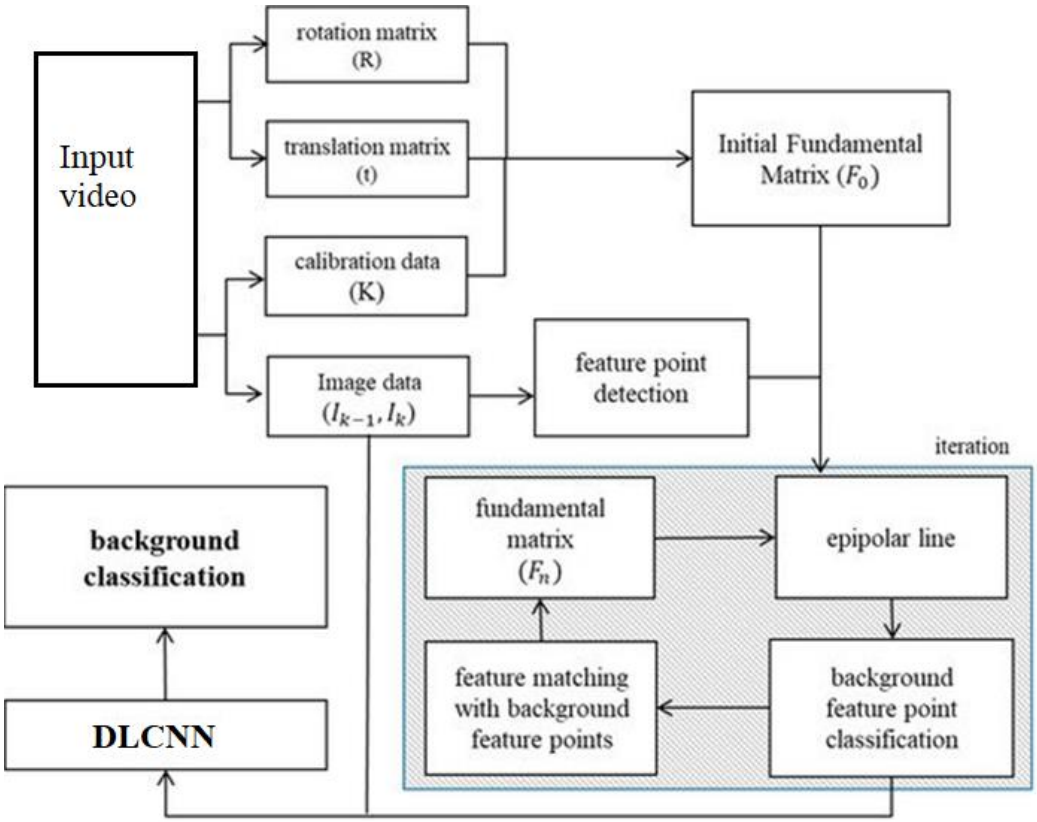


Figure 2: DLCNN based video background and object detection

On the other hand, the epipolar line may be erroneous since the algorithm's correspondences include the moving object's feature points. To produce a more precise basic matrix and epipolar line, we employed both an image and video. The first fundamental matrix of two consecutive frames may be constructed using video data and an image. The epipolar line may be constructed to categories the feature points using this basic matrix. The backdrop feature's matching feature points must sit on the epipolar line in the following frames. As a result, the foreground point is the feature point distant from the epipolar line (i.e., the moving object).

After the background and foreground feature points have been categorized, the fundamental matrix is constructed utilizing the background feature points using a normalized eight-point technique. The epipolar line is then redrawn sans the initial classification's outliers. This classification procedure can be repeated until the basic matrix yields a dependable result. In addition, outliers in the background feature point categorization are removed using the DLCNN algorithm. The suggested categorization technique has the benefit of accurately detecting moving items even in environments where moving objects dominate the frame, which is unlike earlier methods. Furthermore, because the suggested system allows for the complimentary use of the two separate sensors, any video sensor defects are considerably minimized by the image sensor. Figure 2 depicts a high-level overview of the suggested backdrop categorization approach.

The DLCNN technique generates a 128-dimensional descriptor for each feature point, and DLCNN feature points with the shortest Euclidean distance are matched in two adjacent frames. The matching result is shown in Figure 2; we can see that the feature points are concentrated and form tiny clusters. The area around the moving object has a significant number of feature points (called outside points), which is detrimental to image registration accuracy. The following things should be considered while selecting feature points: 1. When the feature points' intensity is higher, they are less likely to be lost; 2. An excessive number of feature points can result in excessive calculations when matching features; 3. An excessive

Published By Scholar Indexing Society
E-ISSN: 2228-837X----DOI URL: http://doi.org/10.53075/ljmsirq/665776577677656
Journal Homepage: https://scholars.originaljournals.com/ojs/index.php/ojs/index

concentration of feature points can result in high motion parameter inaccuracies. As a result, a good feature point distribution should be broad and average, with no outliers.

3.1 DLCNN

DLCNNs are commonly employed to construct motion information since the allocation of spatial attention varies over time and has a certain consistency. This necessitates the use of extra deep convolutional networks, such as DLCNN, which results in an excessive number of model parameters and calculations, which is incompatible with model deployment. Furthermore, because the optical flow network creates local pixels that correlate to motion information, modelling the continuity between high-level semantic properties is problematic. We construct the local attention sequence model, focused on small-range motion information in the local domain, to create the temporal consistency of the spatial attention module, because motion between neighboring moments happens more in the local domain.

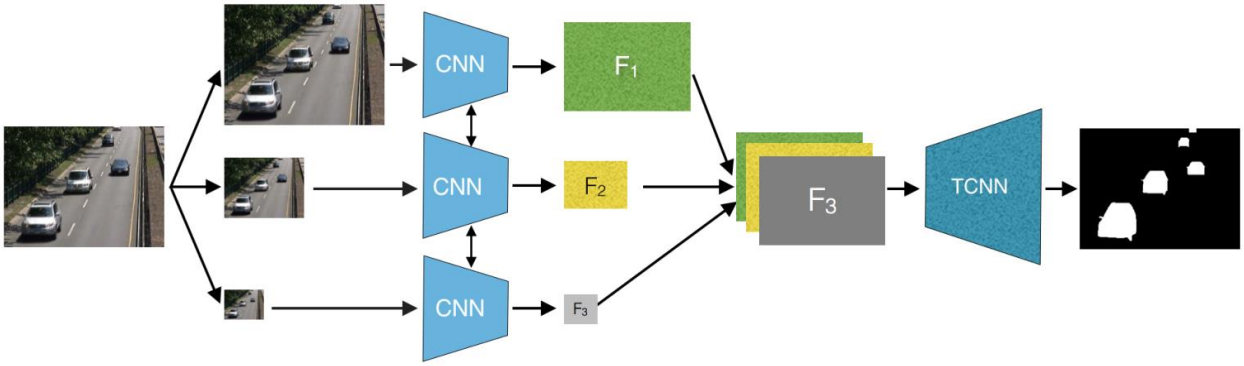


Figure 3: DLCNN object detection

Figure 3 depicts the architecture of the DLCNN technique. The input frame at t is first supplied to DLCNN as the network backbone to extract feature maps. The feature maps are then sent into the event-aware DLCNN, which includes an event detection module to handle difficult events like aspect ratio changes or significant motion. For region proposal creation, the proposed event-aware DLCNN generates improved feature maps. The supporting frame is chosen once all of the ROIs have been created by comparing the feature maps of N preceding candidate frames to the improved feature maps of the current frame. The local attention sequence model can be expressed in the following way: The initial stage is to aggregate the matching feature cells with correspondence weights to obtain an aligned distribution of spatial attention.

The sparse video frame is an extremely sparse image including pixel level video information relating to depth, lateral velocity, and longitudinal velocity. The raw video points are first translated from the video coordinate system to the frame coordinate system using extrinsic calibration parameters to create the sparse video frame. The processed raw video points are then formulated as the sparse video frames utilizing the frame's inherent calibration parameters. The sparse video image is the same size as the frame, and it has three channels for the video characteristics.

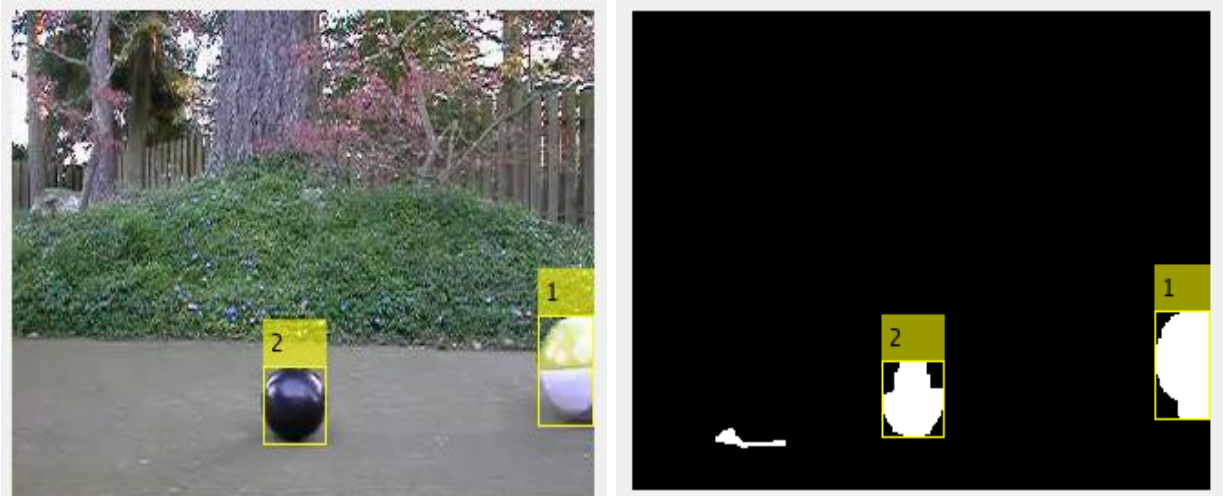
The visible frame feature extraction branch is based on the DLCNN model's pre-trained model. The video feature extraction branch is designed to extract features from the video frame S, which is extremely sparse. Several 2D convolution filters with a single stride are utilized to account for each pixel-level video characteristic. MaxPooling 2D is used to decrease the dimensionality of the feature maps. Output Branch: The features retrieved from the sparse video image and visible frame are transmitted to the output branch using skip connections. In the output branch, the different feature maps are concatenated at various levels. Two sub-branches exist in the output branch based on the small DLCNN model. The first sub-branch detects small and medium-sized impediments, whereas the second sub-branch detects large obstacles. The output branch reshapes layers using numerous 2D convolution filters, which recognised the input video objects.

Published By Scholar Indexing Society
E-ISSN: 2228-837X----DOI URL: http://doi.org/10.53075/ljmsirq/665776577677656
Journal Homepage: https://scholars.originaljournals.com/ojs/index.php/ojs/index

4. Simulation Results:

This section gives a detailed analysis of simulation results, which are implemented by using MATLAB-R2018a software. The DLCNN models are trained with a bulk number of real-time videos with background frames.

4.1 Video object detection performance



(a) (b)

Figure 4: Multi object detection for video sequence 1 (a) original frame (b) background frame

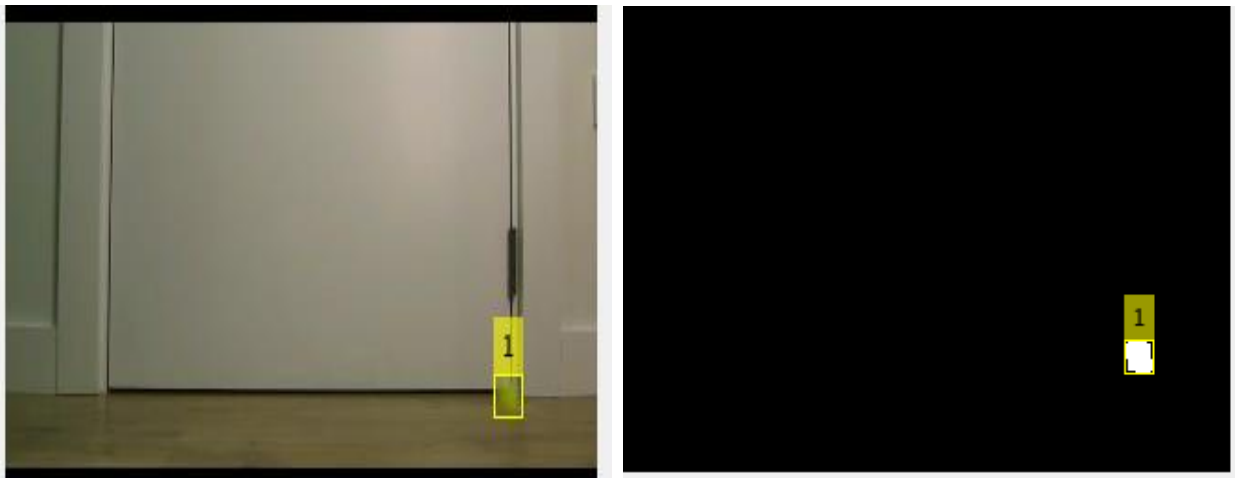


Figure 5: Multi object detection for video sequence 2 (a) original frame (b) background frame

From figure 4 and figure 5, it is observed that the proposed method resulted in the superior performance of object detection. Figures 4(a) and 5 (a) indicate that the proposed method perfectly detected the multiple objects outlined by the yellow color bounding box. Figure 4(b) and figure 5 (b) indicates that the proposed method perfectly detected the multiple objects in the background area also, which is outlined by the yellow color bounding box, respectively.

Published By Scholar Indexing Society
E-ISSN: 2228-837X----DOI URL: http://doi.org/10.53075/ljmsirq/665776577677656
Journal Homepage: https://scholars.originaljournals.com/ojs/index.php/ojs/index

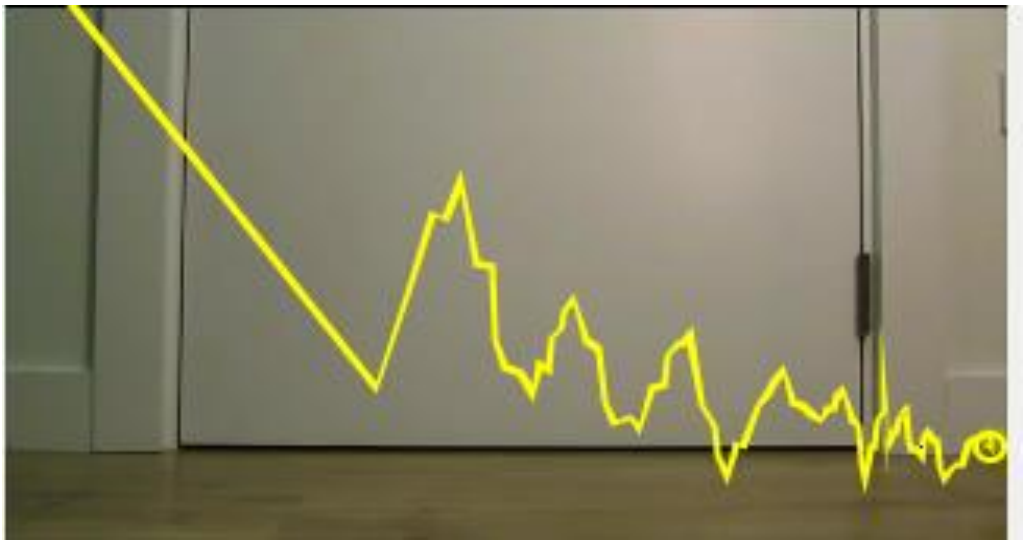


Figure 6: Object Tracking Frame view

Figure 6 shows the object tracking of in the video, the yellow color line indicates the epipolar line and resulted in superior object detection.

4.2 Quantitative evaluation:

Table 1: Performance comparison of video object detection methods

Metric	Sensitivity	Specificity	Accuracy	Precision	F1-score
SVM [11]	93.31	90.48	91.39	91.69	95.34
CNN [13]	94.59	93.49	92.38	94.72	96.37
Proposed	97.l4	95.69	97.36	97.18	98.48

From the table 1, it is observed that the proposed method resulted in optimal performance with respect to the Sensitivity, Specificity, Accuracy, F1-score and precision parameters as compared to the conventional SVM [11] and CNN [13] approaches.

5. Conclusion:

A novel MODT classification approach for video sequences is proposed in this article using DLCNN. The suggested technique increases the accuracy of feature-based object recognition, which has flaws in decoding photos with extensive foreground regions. We used the video sequence to acquire the first fundamental matrix to generate an accurate fundamental matrix. The DLCNN approach is invariant to the foreground section of the frame since it employs a video sequence. The background points are utilized to build a more precise basic matrix after the first categorization of feature points. This procedure can be continued iteratively until the mistake is minimized. The transformation matrix of the subsequent frames is then calculated using the categorized background feature points. This transformation matrix is used to adjust for the backdrop image's mobility. Then, morphological image processing creates the difference map of the two sequential pictures, and the foreground region is retrieved. Finally, the moving object is marked with a minimal bounding box.

References:

[1]. Fulari, Sunit. "A Survey on Motion Models Used for Object Detection in Videos." 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018.
[2]. Maddalena, Lucia, and Alfredo Petrosino. "Background subtraction for moving object detection in RGBD data: A survey." Journal of Imaging 4.5 (2018): 71.

Published By Scholar Indexing Society
E-ISSN: 2228-837X----DOI URL: http://doi.org/10.53075/ljmsirq/665776577677656
Journal Homepage: https://scholars.originaljournals.com/ojs/index.php/ojs/index

[3]. Patil, Prashant W., and Subrahmanyam Murala. "Msfgnet: A novel compact end-to-end deep network for moving object detection." IEEE Transactions on Intelligent Transportation Systems 20.11 (2018): 4066-4077.
[4]. Kalantar, Bahareh, et al. "Multiple moving object detection from UAV videos using trajectories of matched regional adjacency graphs." IEEE Transactions on Geoscience and Remote Sensing 55.9 (2017): 5198-5213.
[5]. Zhang, Junpeng, Xiuping Jia, and Jiankun Hu. "Error bounded foreground and background modeling for moving object detection in satellite videos." IEEE Transactions on Geoscience and Remote Sensing 58.4 (2019): 2659-2669
[6]. Javed, Sajid, et al. "Moving object detection on RGB-D videos using graph regularized spatiotemporal RPCA." International Conference on Image Analysis and Processing. Springer, Cham, 2017.
[7]. Tang, Peng, et al. "Object detection in videos by high quality object linking." IEEE transactions on pattern analysis and machine intelligence 42.5 (2019): 1272-1278.
[8]. Zhang, Junpeng, et al. "Online Structured Sparsity-Based Moving-Object Detection from Satellite Videos." IEEE Transactions on Geoscience and Remote Sensing 58.9 (2020): 6420-6433.
[9]. Teutsch, Michael, Wolfgang Krüger, and Jürgen Beyerer. "Moving object detection in top-view aerial videos improved by image stacking." Optical Engineering 56.8 (2017): 083102.
[10]. Chen, Bo-Hao, Ling-Feng Shi, and Xiao Ke. "A robust moving object detection in multi-scenario big data for video surveillance." IEEE Transactions on Circuits and Systems for Video Technology 29.4 (2018): 982-995.
[11]. Supreeth, H. S. G., and Chandrashekar M. Patil. "Moving object detection and tracking using deep learning neural network and correlation filter." 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2018.
[12]. Mandal, Murari, Lav Kush Kumar, and Santosh Kumar Vipparthi. "MOR-UAV: A benchmark dataset and baselines for moving object recognition in UAV videos." Proceedings of the 28th ACM International Conference on Multimedia. 2020.
[13]. Patil, Prashant W., et al. "Msednet: multi-scale deep saliency learning for moving object detection." 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2018.
[14]. Servin, Martin, et al. "Static and Moving Object Detection and Segmentation in Videos." 2019 Sixth HCT Information Technology Trends (ITT). IEEE, 2019.
[15]. Xu, Yiping, Hongbing Ji, and Wenbo Zhang. "Coarse-to-fine sample-based background subtraction for moving object detection." Optik 207 (2020): 164195.