Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Deep Neural Network-based Multi-Object Tracker in Complex Events

M. Duraipandian

Department of CSE, Nehru Institute of Technology, Coimbatore, India

E-mail: mduraipandiandp@gmail.com

Abstract

Deep learning has been shown to be efficient for multiple object tracking, despite the challenges of frequently occurring occlusions, uncertain appearances, objects in as well as out, and insufficient labeled data. Detecting and tracking objects is one of the most common and difficult jobs that surveillance systems must undertake in order to recognize important events and suspicious conduct, as well as automatically remark and extract video information. The progress of convolutional neural networks (CNN) changes the way objects are tracked. CNN layers trained upon a significant amount of videos or image sequences improve object tracking accuracy in shorter time periods. This study analyses and compares the network model and tracking techniques with its performance measures.

Keywords: Multiple Object Tracking (MOT), Convolutional Neural Networks (CNN), Tracking, Machine Vision, Video Surveillance

1. Introduction

Multiple Object Tracking (MOT), also known as Multi-Target Tracking (MTT), is a computer vision task that analyses videos to detect and track objects with no previous understanding of the appearance or number of targets. There may be categories such as pedestrians, automobiles, animals, and objects that are inanimate. It has long been a significant difficulty in computer vision applications like surveillance, autonomous driving, and human-computer interaction. MOT with track-by-detection must utilize linkage of data among current tracking and unique detections during a single frame in order to build numerous object

DOI: https://doi.org/10.36548/jiip.2023.4.005
Received: 27.10.2023, received in revised form: 19.11.2023, accepted: 12.12.2023, published: 18.12.2023

trajectories. As a result, data association leads to detection sequences with distinct identities. The detection of objects has been quickly progressing in the area of computer vision. Convolutional neural network (CNN)-based models are now achieving remarkable success, particularly in image processing, pattern recognition, and smart surveillance analysis (particularly in object detection, tracking, and recognition). Object detection and tracking are linked together because object detection is the foundation of object tracking, and everyone must pick the proper features and train to achieve successful classification. Person tracking is a difficult issue since people are malleable objects with varying looks, positions, scale, and size.

1.1 Computer Vision on Object Detection and Object Tracker

Object Detection: Object detection is a deep learning system that detects objects in images and videos, such as humans, buildings, and automobiles. The object detection algorithm uses deep learning and machine learning algorithms to provide amazing results.[1]

Machine Learning: Machine learning-based object recognition methods must manually extract features. (Eg: Histogram of oriented gradients (HOG))

Deep Learning: Deep learning-based object detection methods may extract features automatically using deep learning algorithms. (Eg: CNN, Auto Encoder, etc.)

Object Tracker: Deep learning tracking is the challenge of predicting the locations of objects in a video using both spatial and temporal data. Trackers are grouped into various categories, such as tracking techniques and the quantity of items to be tracked.

Single Object Tracker: Even when there are several other things on the screen, these trackers only track a single object. They function by initially initialising the object's position in the initial frame, followed by tracking them throughout the series of frames. These tracking technologies are quite rapid [3].

Multi-Object Tracking: Tracking numerous items entails tracking many objects. The tracking techniques must first compute the number of objects within each frame and keep each object's identifier from frame-to-frame. Tracking the state of an unknown and changing number of items by calculating noisy data is known as multi-object tracking (MOT), and it has

significant applications including self-driving cars, tracking animal behavior, defence systems, and others [4].

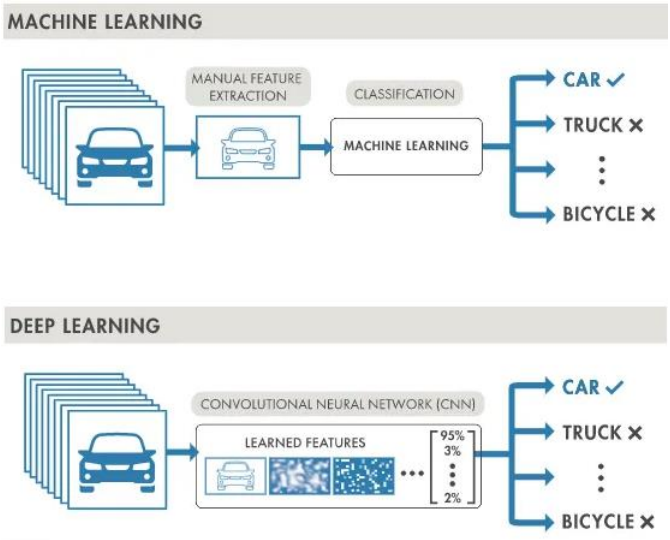

Figure 1. Machine Learning and Deep Learning based Object Detection [1]

It is mainly classified into three categories in Open CV: Image classification, Object Localization, Object detection.

1.2 Advantages of Object Tracker from Object Detector

- **Tracking Helps with Detection Fails**: There are several scenarios in which an object detector may fail. However, if there's an object tracker within place, it may still forecast the items in the frame.
- **Assignment of ID**: When we use a detector, it just shows us where the items are; if we simply glance at the array of outputs, we won't know which coordinates correspond to which box. A tracker, allocates ID to tracked items as well as holds that ID in the frame of the object's life span.
- **Implementation in the Real-Time Scenario:** Trackers are extremely quick and, in general, outperform detectors. Trackers may be employed in real-time settings because of this characteristic.

2. Related Works

[5] In this individual have been identified in above projection video clips and are tracked using GOTURN algorithm. The faster-RCNN detection model achieved a true detection rate of 90%- 93% having 0.5% false rate. The GOTURN tracker method yielded similar results, with rates of success 90-94%.

[6] Avşar et al., proposes a Vehicle detection by YOLOv4, whereas vehicle tracking is handled by either the Kalman filter or the Deep SORT algorithm throughout the video. In a 20-minute video comprising 297 automobiles, the suggested technique achieves an absolute error of 14 vehicles, corresponding to a normalized absolute error rate of 1.571%. The identical error metrics calculated with the benchmark approach are 33 cars and 3.704% in the same video. When Deep SORT is utilised as the vehicle tracker, error rates are reduced.

[7] Jo et al., YOLOv2, a quick and accurate DL based method, is combined to fast tracker technique to create real-time a MOT system of high accuracy. YOLOv2, the Kalman filter, and the Hungarian algorithm are used to monitor multi-class multi-objects in real-time. Also, TUD-crossing and ETH-crossing datasets are subjected to multi-object pedestrian tracking algorithms based on multi-class multi-object tracking algorithms.

[8]Jiang et al., proposes an unique strategy for multi-object tracking using multi-agent deep reinforcement learning (MADRL) to address issues with existing MOT methods like variable number of targets, non-causality, as well as non-realtime. YOLO V3 is used to detect the objects within the frame.

[9] Baisa et al., proposes a visual MOT constructed using a Gaussian mixture Probability Hypothesis Density (GM-PHD) filter and a similarity Convolutional Neural Network (CNN). The study uses spatiotemporal and visual similarities derived from object bounding boxes as well as deep CNN appearance characteristics, respectively, to compensate for the study's lack of tagging targets across frames.

[10] Ahn et al., proposes a Faster region-based convolutional neural network (R-CNN) deep learning method with geometric changes in conjunction with multi-camera or multi-image fusion technology that is used in this multi-object tracker for mice. The technology monitored every individual within groups of unlabeled mice and was used to study pursuing behaviour.

[11]Xia et al., Initially, CNN is used for identifying and determining the target bounding box. Second, to achieve the preliminary multi-object tracking, we employ a particle filter (PF) as the tracker. Finally, mutual management of the PF tracker and CNN detector produces the multi-object tracking trajectory. We utilise the forward-backward (FB) distortion of the tracker at a certain point to evaluate the outcomes of the experiment.

[12] Zhang et al. proposed an aYOLOv3-DeepSort-based processing framework for analyzing pedestrian-vehicle interaction behavior. On the MOT16 dataset, YOLOv3-DeepSort can attain ML values of 14.10% and ID values of 382.

More research that focuses on multi-object detection is given below

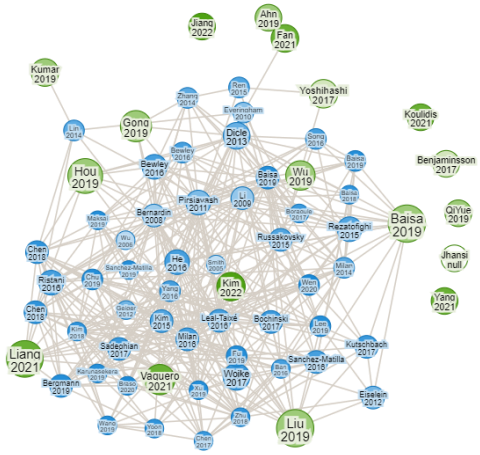

Figure 2. Evolution of Related Works with Reference to MOT

3. Multi-Object Detection Algorithms

It refers to the process of categorizing and identifying the location of an item in an image. This phase must be completed prior to object tracking since we need to determine where a particular object exists at each frame. It is concerned with localizing a specific ROI of an image and categorizing is as an image classifier does. A single photograph might have many areas that are interesting, pointing to various items. Object identification thus becomes a more complicated image classification task. It is essential to many applications, including surveillance, self-driving automobiles, and robots.

There are two types of object detection algorithms:

- Single-Shot Detectors
- Two-Stage Detectors.

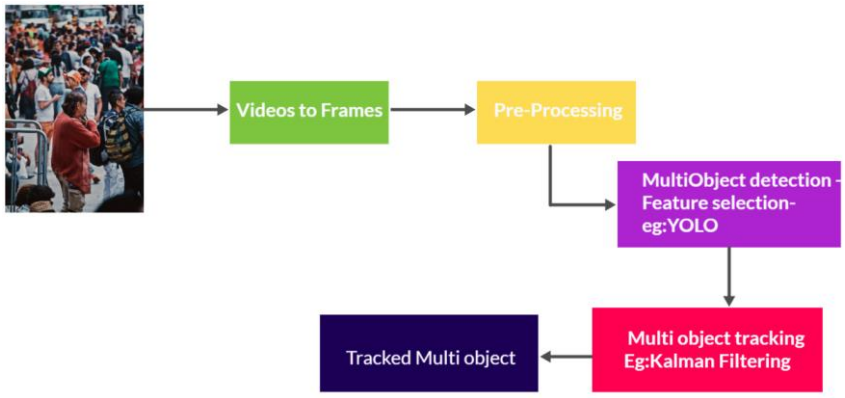

Figure 3. Process Flow of The Multi-Object Detection and Tracking

Single-shot object detection makes predictions regarding the existence and positioning of objects within the image by utilizing just one pass through the input image. It processes a complete image in a single pass, resulting in increased computing efficiency. Single-shot object detection is less accurate than other approaches in general, and it is less successful at recognizing small objects. In resource-constrained contexts, such methods can be utilized to identify objects in real-time. YOLO is a single-shot detector that processes an image using a fully convolutional neural network (CNN).

3.1 YOLO V3

YOLO considers the detection issue to be a regression problem in order to obtain direct predictions of probability classes and bounding boxes using features of an input image. The algorithm of YOLO is:

1. The input image (Im) is split into grids (g x g).
2. Object classification is performed using extracted features from every grid cell.
3. Each of the grid cells predicts the B bounding boxes.
4. C: class probabilities and C: object classes are evaluated for each bounding box.

5. For each bounding box, two measurements are taken into account.

    - Bounding box probability (P) is defined to determine whether or not the bounding box corresponds to any object.

    - The IoU between the ground truth and the bounding-box is defined in order to determine how precisely the bounding-box includes that object.

6. The object area is defined as the bounding box with the highest IoU and a non-zero class probability.

The YOLOV3 detector is only capable of single-class prediction with high-resolution detection.

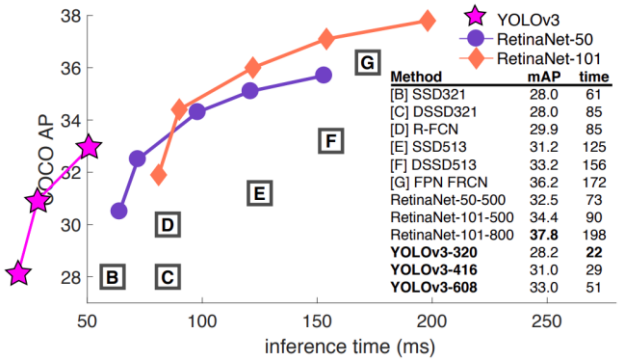

Figure 4. Comparison of Yolo and its Inference time taken for Detection [13]

3.2 Two-Stage Detectors: Faster RCNN

The most frequently used two-stage detector is Faster RCNN. R-CNN added a Region Proposal Network (RPN), which made it significantly quicker than earlier versions (R-CNN and Fasy RCNN) . These models are well-known for their speed and high mean average precision (MAP) in object recognition. The RPN operates by sliding a window over the feature map, creating k squares of fixed size, known as anchor squares, of various sizes and shapes in each one. A grouping layer is used to resize the collection of regions of interest (ROI) without an associated class to a fixed size appropriate for the network after the ROI is projected onto the maps of the convolutional characteristic using the RPN.

3.3 Multi-Object Tracker Steps

The objective of object tracking is to estimate the bounding boxes and IDs of objects in motion pictures. It accepts a series of initial item detections, creates a visual model for the objects, and then tracks them as they move about in a movie.

SORT promotes real-time performance and incorporates a Kalman filter with the Hungarian algorithm as tracking components.

Step 1: The tracking module begins at this step. An object detector identifies the items to be tracked in the frame during this stage. These detections are subsequently forwarded to the next stage.

Step 2: In this stage, detections are transmitted from one frame to the next, calculating the position of the target's frames using the velocity constant model. If detection is linked to the frame of the target, using bounding box, the detected is used for target state modification, with velocity components addressed by Kalman filter model.

Step 3: We now have both the target and detected bounding boxes. As a result, the intersection-over-union (IOU) distances between each detection and all anticipated bounding boxes of the existing targets is generated as a cost matrix.

Step 4: This module is in charge of ID generation and deletion. According to the IOUmin, unique identities are produced and discarded. If the overlap between detection along with target is smaller than IOUmin, the object is untracked.

3.4 DeepSORT

Despite having overall strong tracking accuracy and precision, SORT has a large number of identity switches. It fails in numerous difficult conditions, such as occlusions, various camera angles, and so forth. DeepSORT employs a more effective association measure that incorporates both motion along with appearance descriptions. DeepSORT is a tracking algorithm that monitors objects not just based on their velocity and motion, but also on their appearance. For that, the characteristics are distinguished before applying tracking, and embedding is trained offline. The network was trained using a large-scale human re-

identification dataset, making it suited for tracking context. The DeepSORT cosine metric method of learning is used for training the deep associate metric model.

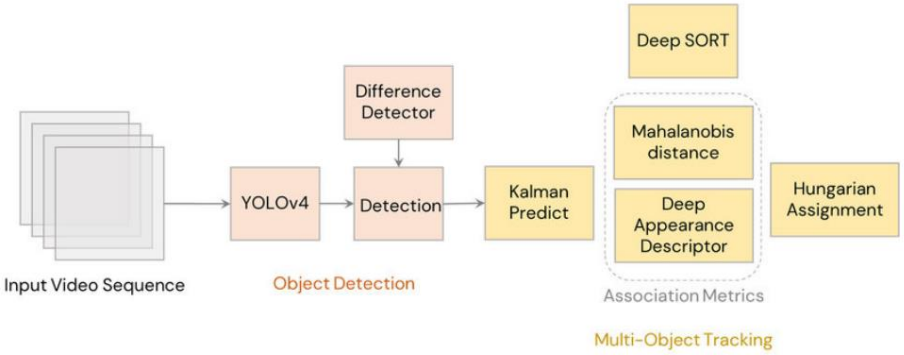

Figure 5. Multi-Object Tracking Process Flow [14]

3.5 **Comparison Analysis**

The below table shows the comparative analysis of the multi-object tracking algorithms and their performance metrics.

Table 1. Comparison of Different Implemented Deep Learning Techniques

Related articles	**Method**	**Dataset**	**Description**
15]	DeepSORT_Y+RN*	VisDrone 2018	MOTA ↑-45.8 MOTP ↑0.219
[16]	SORT	MOT17	44.93 87.84
[17]	YOLOv7+ BoT-SORT	MOT20	Accuracy - 69.36%,
	SORT-C3D	Jilin-1 satellite constellation	improve vehicle tracking performance in satellite video
[18]	Deep SORT+YOLO	UA-DETRAC dataset	7 instances, 11.025 frames per second, and 25.193

			bounding boxes were collected.
[19]	CNN+ spatial-temporal attention mechanism (STAM)	MOT15 and MOT16	MOTA -MOT accuracy-34.3% on MOT15 and 46.0% on MOT16
[20]	Faster R-CNN	MOT15	MOTA -53.0 MOTP- 75.5
[21]	CNN- FANTrack	KITTI	MOTA - 77.72 MOTP-82.23
[22]	Single shot detector +Embedded learning	MOT-16 challenge	MOTA-64.4-for 22 to 40 FPS
[23]	MultiSSD + CCF (CorrelationFilter tracker)+deep CNN	MOT2015	MOTA 32.7%
[24]	RNN- Bilinear LSTM	MOT 2017	MOTA- 47.5

4. Challenges in Multi-Object Tracking

- Object tracking is difficult in the real world because several factors will affect the object of interest. Being aware of these typical concerns is the first step in addressing them while building object-tracking algorithms.
- Variations occur as a result of geometric alterations: scale of the object, pose , etc.
- The target object's IDs get swapped owing to identical attributes such as similar attire, facial structure, spectacles, skin colour, height, and so on.
- When the background is heavily occupied it is difficult to extract features, identify or even track a target item because it contributes additional redundant data or noise, making the network less sensitive to key characteristics.

- Drifting as a result of an incorrect target model update. One incorrect update may cause a constant updating in the wrong direction, causing the viewer to forget the right direction throughout the video.

- Based upon the resolution, the total number of pixels within the dataset used for training bounding box may be too low for accurate object tracking.

5. Conclusion

In this research, the study presents a dynamic CNN-based online MOT method that efficiently exploits the advantages of multi-object trackers by using CNN features. MOT research benchmark datasets and evaluation techniques are analyzed. The study is looking for a fundamental development in the application of multi-object tracking techniques to real-world problems. To address challenges like frequent IDs and incorrect dense multi-object tracking, the study analysis innovative network models, training methods, performance measures and so on. Finally, this survey examined the main challenges of MOT and suggested potential solutions that can be researched further.

References:

[1] https://medium.com/visionwizard/object-tracking-675d7a33e687

[2] https://microsoft.github.io/computervision-recipes/scenarios/tracking/FAQ.html

[3] https://encord.com/blog/object-tracking-guide/

[4] Pinto, Juliano, Georg Hess, William Ljungbergh, Yuxuan Xia, Henk Wymeersch, and Lennart Svensson. "Can deep learning be applied to model-based multi-object tracking?." *arXiv preprint arXiv:2202.07909* (2022).

[5] Jiang, Ming-xin, Chao Deng, Zhi-geng Pan, Lan-fang Wang, and Xing Sun. "Multiobject tracking in videos based on lstm and deep reinforcement learning." Complexity 2018 (2018): 1-12.

[6] Avşar, Ercan and Yağmur Özinal Avşar. “Moving vehicle detection and tracking at roundabouts using deep learning with trajectory union.” *Multimedia Tools and Applications* (2022): 1-28.

[7] Jo, KangUn, Jung-Hui Im, Jingu Kim and Dae-Shik Kim. “A real-time multi-class multi-object tracker using YOLOv2.” *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* (2017): 507-511.

[8] Jiang, Mingxin, Tao Hai, Zhi-geng Pan, Haiyan Wang, Yinjie Jia and Chao Deng. “Multi-Agent Deep Reinforcement Learning for Multi-Object Tracker.” *IEEE Access* 7 (2019): 32400-32407.

[9] Baisa, Nathanael L.. “Online Multi-object Visual Tracking using a GM-PHD Filter with Deep Appearance Learning.” *2019 22th International Conference on Information Fusion (FUSION)* (2019): 1-8.

[10] Ahn, Hyochang and Han-Jin Cho. “Research of multi-object detection and tracking using machine learning based on knowledge for video surveillance system.” *Personal and Ubiquitous Computing* 26 (2019): 385-394.

[11] Xia, Yu, Shiru Qu, Sotirios Goudos, Yu Bai, and Shaohua Wan. "Multi-object tracking by mutual supervision of CNN and particle filter." Personal and Ubiquitous Computing (2021): 1-10.

[12] Zhang, Qiang. "Multi-object trajectory extraction based on YOLOv3-DeepSort for pedestrian-vehicle interaction behavior analysis at non-signalized intersections." Multimedia Tools and Applications 82, no. 10 (2023): 15223-15245.

[13] Redmon, J. "YOLOv3: An Incremental Improvement/Joseph Redmon, Ali Farhadi-University of Washington." (2018).

https://www.v7labs.com/blog/yolo-object detection#:~:text=a%20negative%20prediction.-,What%20is%20YOLO%3F,repurposed%20classifiers%20to%20perform%20detection.

[14] https://learnopencv.com/understanding-multiple-object-tracking-using-deepsort/#Simple-Online-Realtime-Tracking-(SORT)

[15] Kapania, Shivani, Dharmender Saini, Sachin Goyal, Narina Thakur, Rachna Jain, and Preeti Nagrath. "Multi object tracking with UAVs using deep SORT and YOLOv3 RetinaNet detection framework." In Proceedings of the 1st ACM Workshop on Autonomous and Intelligent Mobile Systems, pp. 1-6. 2020.

[16] Pereira, Ricardo, Guilherme Carvalho, Luís Garrote, and Urbano J. Nunes. "Sort and deep-SORT based multi-object tracking for mobile robotics: Evaluation with new data association metrics." Applied Sciences 12, no. 3 (2022): 1319.

[17] Li, Tingting, Zhanbo Li, Yuhong Mu, and Jie Su. "Pedestrian multi-object tracking based on YOLOv7 and BoT-SORT." In Third International Conference on Computer Vision and Pattern Analysis (ICCPA 2023), vol. 12754, pp. 369-374. SPIE, 2023.

[18] Meimetis, Dimitrios, Ioannis Daramouskas, Isidoros Perikos, and Ioannis Hatzilygeroudis. "Real-time multiple object tracking using deep learning methods." Neural Computing and Applications 35, no. 1 (2023): 89-118.

[19] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, Nenghai Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4836-4845

[20] Chen, Long, Haizhou Ai, Chong Shang, Zijie Zhuang, and Bo Bai. "Online multi-object tracking with convolutional neural networks." In 2017 IEEE international conference on image processing (ICIP), pp. 645-649. IEEE, 2017.

[21] Baser, Erkan, Venkateshwaran Balasubramanian, Prarthana Bhattacharyya, and Krzysztof Czarnecki. "Fantrack: 3d multi-object tracking with feature association network." In 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 1426-1433. IEEE, 2019.

[22] Wang, Zhongdao, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. "Towards real-time multi-object tracking." In European Conference on Computer Vision, pp. 107-122. Cham: Springer International Publishing, 2020.

[23] Zhao, Dawei, Hao Fu, Liang Xiao, Tao Wu, and Bin Dai. "Multi-object tracking with correlation filter for autonomous vehicle." Sensors 18, no. 7 (2018): 2004.

[24] Kim, Chanho, Fuxin Li, and James M. Rehg. "Multi-object tracking with neural gating using bilinear lstm." In Proceedings of the European conference on computer vision (ECCV), pp. 200-215. 2018.