

# A cost-effective, reliable and accurate framework for multiple-target tracking by detection approach using deep neural network

Divyaprabha<sup>1</sup>, Guruprasad Seebaiah<sup>2</sup>

<sup>1</sup>Department of Electronics and Communication Engineering, Sri Siddhartha Institute of Technology, Tumkur, India

<sup>2</sup>Department of Biomedical Engineering, Sri Siddhartha Institute of Technology, Tumkur, India

## Article Info

### Article history:

Received Feb 27, 2024

Revised Jul 5, 2024

Accepted Jul 9, 2024

### Keywords:

Computer vision

Convolutional neural networks

Data association

Deep learning

Multiple target tracking

PersonRe-IDNetwork

## ABSTRACT

Over the years the area of object tracking and detection has emerged and become ubiquitous owing to its potential contribution towards video surveillance applications. Multiple object tracking (MOT) estimates the trajectory of several objects of interest simultaneously over time in a series of video frames. Even though various research proposals have encouraged the use of machine learning techniques in designing multi-object trackers, the existing solutions need to be more practicable for online tracking due to more complicated algorithms. The study, therefore, introduces a cost-effective tracking solution for multiple-target tracking by detection where it incorporates the you only look once version 4 (YOLOv4) and person re-identification network, which are further integrated with the proposed tracking model, which considers both bounding box and appearance features to handle the motion prediction and data association respectively. The novelty of this approach lies in considering appearance features, which not only help predict tracks through allocations problem solving but also handle the cost of computation problems. Here, the system utilizes a pre-trained association metric with which the occlusion challenges are also handled, whereas the target tracking has taken place even in more extended periods of occlusion, making it suitable with the existing efficient tracking algorithms.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Divyaprabha

Department of Electronics and Communication Engineering, Sri Siddhartha Institute of Technology

839V+8G9, Tumkur-Kunigal Rd, Saraswathipuram, Tumkur, India

Email: divyasy74@gmail.com

## 1. INTRODUCTION

Over the years, the underlying concept of object detection and tracking has emerged and is considered one of the most essential and challenging branches of computer vision. However, its ideas have been widely applied in various fields, including healthcare monitoring, autonomous driving, anomaly detection, vehicle detection and pedestrian tracking [1]. It can also be seen that a variety of modalities, including radar, light detection and ranging (LIDAR), and computer vision (CV), have become available for the purpose of object detection and tracking. In recent years, significant progress could be seen even in the field of imaging technology as cameras are now cheaper, smaller and of higher quality than ever before. On the other hand, computing platforms have been optimized toward parallelization, such as multi-core processing and graphical processing units (GPUs). Such versions allow CV towards efficient object detection and tracking to pursue real-time implementation. The rapid development of deep learning (DL) networks along with GPU's computing power has significantly contributed towards improving the performance of

object detectors and trackers. The concepts behind deep convolutional neural network (CNN) and GPU's enhanced computing power are the prime reasons behind fast evolution of CV-based object detection and tracking [2]. In this context it has to be mentioned that deep learning ideas have evolved from machine learning (ML) and also there comes characteristics differences. Basically, ML is considered as a branch of artificial intelligence (AI) where it basically learns patterns from examples or sample data. Here the machine is designed in such a way that it has the ability to access and learn from the data. Here the data could be labeled, unlabeled or their combination. The learning could be either supervised, un-supervised or semi-supervised. Artificial neural networks (ANNs) are subjected to have the ability to learn the relation between input and output from examples and are basically considered as potential solutions of ML. In the early 2000s, certain breakthroughs in engineering research for multi-layered neural networks (MLP) had aid in arrival of deep learning. DL refers to a learning in depth paradigm which involves different layers and stages [3]. Unlike conventional shallow learning models, DL is dependent over hundreds or thousands of frames as a result it has become computationally intensive and difficult to engineer which is considered as a prominent research challenge by studies [4] and [5]. It requires a very high-performance GPU to provide fast and accurate object recognition and motion detection.

Even though DL models are mostly complex in design and having dependency over huge amount of training data and GPU's computing power but these factors can be better handled with DL models as claimed by [5] and [6]. It has been found that there exist various benchmark datasets for object detection such as California Institute of Technology (Caltech) [6], Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) [7], ImageNet [8], PASCAL visual object classes (PASCAL VOC) [9], Microsoft Common Objects in Context (MS COCO) [10] and The Open Images Dataset V5 [11]. Also due to the availability of huge amounts of datasets and the rapid development of GPUs, DL models are widely adopted for object detection by researchers and the advancement is still in progress.

Bochkovskiy *et al.* [12] addresses the design challenges of two stage detectors such as region convolutional neural network (RCNN) [13] and fast RCNN [13] which is the next advanced version of RCNN. RCNN is considered to be the first model of two stage detection which shows that deep-CNN is better than convolutional method for object detection. On the other hand, RNN is found to be involved in multi-stage training process where as fast RCNN use only one stage end-to-end training process. Hussain *et al.* [14] in their study highlights various deep learning-based solutions for detecting objects in image frames or videos using several algorithms which include you only look once (YOLO), spatial pyramid pooling (SPP-net), fast R-CNN, CNN based on regions (R-CNN), orient gradient histograms (HOG), single shot detector (SSD), region-based fully convolutional network (R-FCN), however the study made a claim that even though object detection algorithms are found to be accurate from the detection accuracy point of view but the speed of inferencing is always being a problem. YOLO has become very popular in object detection and considered to be state-of-the-art technology in object detection algorithm. YOLO is found to be more accurate and faster when compared with RCNN and fast RCNN and also found suitable for dynamic multi-object tracking. In the context of YOLO, object detection is considered to be the regression problem. It has to be noted that this approach is built on the top of CNN found suitable for real-time object detection. In the context of object detection one stage detectors have become popular which include YOLO [15], YOLOv2 [16], YOLOv3 [17]. Here YOLO stands for you only look once. However, it has been observed in the existing system that even though compared to RCNN, YOLO reduces the background false positives by three times, but it has limitations factors such as it imposes strong spatial constraints on bounding box predictions. A significant improvement over YOLO has taken place in its second iteration which is namely YOLOv2 which has improved the speed and the detection precision to some extent. But its limitation comes with high resolution detection and single class objects which is further addresses in YOLOv3 which is the advanced version of YOLOv2, and it incorporates deep CNN DarkNet-53 for the feature generation modeling. It is basically capable of multi-class classification and is used for mostly small object classification problem, however, YOLOv3 model lacks effectiveness for the detection of medium and large sized objects for which limits its scope of applicability in real-time tracking operations. However, Wang *et al.* [18] exhibits the advantage of YOLOv4 in multi object detection and proposes a new variant of YOLOv4 through which it targeted to enhance the multi-object tracking performance as well.

However, the baseline approach of Liu *et al.* [19] is also considered in the proposed study where it is found that the data association and track management policies in recent algorithms faces difficulties in dealing with motion prediction and occlusion handling and also object person re-identification entification. Many recent algorithms are found to employ motion and appearance indications to deal with such challenges but end up increasing the computation cost and for this reason the speed of the algorithm also decreases which makes them unsuitable for online tracking applications. Also there exist various algorithms which only uses motion cues in contrast to increase the speed of algorithm but cannot properly handle occlusions and person re-identification identify the lost objects as shown in the studies by [20]–[22].

Discussion on related research exhibits that significant progress has been made in the line of research for object detection and tracking-by-detection approaches, which have also significantly contributed towards multiple object tracking. However, the challenge arises to appropriately deal with object trajectories, mostly considered global optimization problems. Related frameworks such as the idea of global data association for multi object tracking (MOT) where flow network formulations could be found, also globally optimal greedy algorithms are used for tracking variable number of objects. The use of probabilistic graphical models mainly processes the video batches simultaneously where non-linear motion patterns along with robust appearance features are explored and evaluated. However, most of these approaches of MOT are computationally expensive and do not provide scope of applicability in online scenarios where target identities could be available in each time step. The proposed addresses these challenges and further introduces a cost-effective, reliable and accurate MOT design framework considering YOLOv4 and association metric for the purpose of tracking and discriminating pedestrian entities movement across a street. In this joint approach of MOT, YOLOv4 contributes towards obtaining the detections of objects in individual frames. And also, further through connotation metric the proposed study model executes track of association and management across the frames to determine the tracking and discriminating object of interest movement which is pedestrian in the proposed study context.

## 2. METHOD

The prime objective of this study is to introduce a joint framework of MOT which explores the strength factors of association metric by means of introducing single hypothesis tracking method with Kalman filtering (KF) approach. Here the design and development of the framework for multi-object of interest tracking aims to offer cost-effectiveness along with accuracy in discriminating the pedestrian movement patterns. Here the design and modeling of the tracker will plan to utilize the deep appearance features from the deep neural network (DNN)-based model to improve the tracking accuracy of identification of pedestrians' entities. The prime objective of this proposed study model will be to estimate the count of object of interest in a given video scene. It will also aim to estimate their position precisely while maintaining their unique form of identities. The study model here aims to perform multi-object tracking even in the presence of over occlusions. Here the work-flow strategic design of multi-object tracking will also consider efficient and simplified execution steps so that cost of operations can be controlled with appropriate data association metrics. The study also aims to implement the DNN model for the purpose of allocations operations and also predict the tracks using KF methods. The study also here will introduce a novel cost model to optimize the cost of overall allocations operations for both detections and tracks. The study further elaborates the methodology adopted in the proposed system for MOT.

### 2.1. YOLO object detector

YOLO is considered to be an efficient object detector with one stage which was designed after Faster RCNN. The advanced variant of YOLOv3 considers multi-level classification and adopts more complex datasets which might contain overlapping labels. It utilizes three different scale feature maps to predict the bounding box. In YOLOv3 the last convolutional layer integrates 3D tensor encoding class predictions, objectness and bounding box. However, it is observed that YOLOv3 introduces more effective and robust feature extractor where inclusion of DarkNet-53 can be seen in [23], [24]. The proposed study for designing cost-effective, reliable and accurate framework of MOT considers the potential aspects of different variants of YOLO as baseline models and finds the strength of YOLOv4 as a state-of-art algorithm which is 12% faster compared to the previous version of YOLOv3. It also realizes that if an effective tracking framework model is designed effectively considering YOLOv4 detection then it can significantly contribute towards MOT which could be comparable with RCNN and Faster RCNN.

The study considers YOLO one-stage detector as baseline reference and explores its advantage factors to build the proposed tracking solution. The conventional YOLO network basically consists of 24 convolution layers and 2 fully connected layers. In its advance version of YOLOv2, the improvement can be seen over speed and detection accuracy where major tasks include six major computational steps viz. i) batch normalization, ii) high resolution classifier, iii) convolution with anchor boxes, iv) size and aspect ratio prediction of anchor box, v) analysis of fine-grained features and vi) multi-scale training phase [23].

### 2.2. YOLOv4

The baseline design idea of YOLOv4 is a one-stage detection network modeling basically consists of three core segments which include backbone, neck and head. In the design of the proposed strategy of MOT, the backbone of YOLOv4 can be a pre-trained CNN such as VGG16 or CSPDarkNet53 which trained on the respective datasets. Here in the proposed study context the backbone part of YOLOv4 model enables

feature extraction network which is suitable for object detection. In feature extraction stage the study model computes feature maps ( $f_{map_{l_n}}$ ) which is obtained from the input image frames of pedestrian datasets.

The YOLOv4 network incorporates CSPDarkNet53 module as the backbone which performs feature extraction from the input image frames. The backbone basically connects with five residual layers. The proposed MOT framework also integrates YOLOv4 spatial pyramid pooling (SPP) layer [25] and path aggregation network (PAN) [26]. Here SPP in neck part basically enables concatenation of max-pooling outputs. Here the max-pooling outputs are constructed in the form of low-resolution feature map to not only extract the most representative features, but it also helps in feature dimensionality reduction from computing point of view. The backbone CSPDarkNet53 is designed based on core cross-stage partial network features. Here the output of  $m$ -layer CNN be expressed as (1):

$$y = F(x_0) = x_k = H_k(x_{k-1}), H_{k-1}(x_{k-2}), H_{k-2}(x_{k-3}), \dots \dots H_1(x_0), x_0 \quad (1)$$

Here in (1)  $F(\cdot)$  denotes the mapping function whereas  $x_0$  denotes the input to target  $y$ . Here  $H_k$  is an operational function for the  $k$ th layer of the CNN. However, the core emphasize is to optimize the function  $H_i$  at different layer. It can be achieved with the expression (2).

$$y = M([x_0', T(F(x_0''))]) \quad (2)$$

Here for optimization modeling  $x_0$  is represented into two distinct partitions which can be shown as  $x_0 = [x_0', x_0'']$ . Here  $T$  is the transition function to truncate the gradient flows of  $H_i$  whereas  $M$  is another transition function which is used to mix two segmented parts. However, in the proposed study context the YOLOv4 backbone consists of five residual blocks and output from the residual blocks are aggregated at the neck of YOLOv4 network. The design of SPP layer [1] in the proposed study follows a configuration setting where it incorporates kernels of size  $1 \times 1$ ,  $5 \times 5$ ,  $9 \times 9$ , and  $13 \times 13$  which are subjected to perform max-pooling operations where the cross-stage partial network also contributes towards improving the learn ability and parameter utilization. Here in the proposed study context the stride value is set to 1. The system framework integration for multi-target tracking also configures the YOLOv4 network in such a way where it concatenates the ( $f_{map_{l_n}}$ ) considering the SPP pooling operations. Here the advantage of employing concatenation function is that it increases the receptive field of backbone feature components and also increases the accuracy of YOLOv4 network towards detection of smaller objects. The concatenated feature maps generated from the YOLOv4 SPP module which is represented with  $f_{map_{SPP}}$  are further fused with the high-resolution feature maps  $f_{map_{HR}}$  using PAN as (3):

$$f_{map_{comb}} \leftarrow f_{map_{SPP}} \oplus f_{map_{HR}} \quad (3)$$

Here the PAN component basically employs up sampling and downsampling operations to set bottom-up and top-down paths which further concatenate low-level and high-level features. Further the PAN segment enables the output of aggregated  $f_{map_{comb}}$  which is further used for prediction. Here in the proposed stage the YOLOv4 basically consists of three different prediction heads which are YOLOv3 networks that further generate the final predictions. Here the YOLOv4 network outputs the  $f_{map}$  of sizes  $19 \times 19$ ,  $38 \times 38$ , and  $76 \times 76$  for the purpose of predictive analytics.

### 2.3. Multi-target tracking strategy

In the proposed concept of multi-object or target tracking policy integrates the YOLOv4 for the purpose of precise object detection. Further it aims to perform multi-target tracking even in the presence of occlusions and yet accomplishes higher computational efficiency from the perspective of execution. The proposed context of tracking-by-detection approach initially obtains the detections of objects in each frame and further it tracks the association and management across the video frame sequences. The study addresses the computational complexity challenges in existing approaches and also aims to build an efficient tracking model that can easily deal with the occlusions. In this regard along with YOLOv4, it also integrates the person re-identification networks towards computing the appearance features from the video frames. Here the appearance features ( $F_A$ ) are also called as appearance embeddings ( $E_A$ ) which basically represents the visual appearance of an object within the frame. Here the importance of  $F_A$  is that it offers an additional measure of similarity factor (or distance) between the detections and track or line of movement (LoM). Here the proposed strategy of multi-object of interest tracking utilizes the association metric concept where it also

adopts the single hypothesis tracking methodology using KF approach [27]. Here the study also integrates the appearance information into the frame-by-frame data association to handle the tracking over occlusion. This approach also minimizes the number of switches in track identities. The proposed framework of multi target tracking approach basically relies on three different types of distances for allocations stage which are bounding box IoU, Mahalanobis distance, and appearance cosine distance. In the proposed multi-target tracking model, the study basically formulates two distinct branches which are appearance branch and motion branch respectively. Here the proposed study initially formulates a functional block for track handling and state approximation. The proposed strategy enables the tracker with different features viz. i) estimation filter, ii) association cost, iii) association type, and iv) track maintenance.

#### 2.4. Inclusion of estimation filter and track association with measures

The proposed strategy initially defines KF by means of introducing a detection measurement which is represented with a 2D bounding box. Here the detection measure net for KF can be represented with the (4).

$$Z = [x, y, w, h] \quad (4)$$

Here  $Z$  denotes the detection measurement for 2D bounding box. Also, here  $(x, y)$  represents the coordinates in the form of pixels correspond to the top-left corner of bounding box whereas  $(w, h)$  represents the width and height of the bounding box. However, defining the state estimation filter for eight-dimensional space using a standard KF for bounding box tracking can be represented as  $(u, v, r, h, \dot{u}, \dot{v}, \dot{r}, \dot{h})$ . Here  $(u, v)$  represents the bounding box position in center,  $r$  represents aspect ratio and  $h$  represents height along with their respective velocities in image coordinates. However, in standard KF estimation model the bounding coordinates such as  $(u, v, r, h)$  are taken as a direct observation of the object state. However, in the proposed strategy the estimation of state of estimated bounding box can be measured with the following mathematical expression.

$$X = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}, \dot{r}] \quad (5)$$

Here  $s$  represents the scale of bounding box and  $r$  implies width to height ratio associated with the bounding box. However, unlike the existing KF based state estimation strategy, here the state estimation measurement in conversion can be represented with the time rate of change associated with the aspect ratio. The following mathematical expressions are formulated to perform conversion of measurement to state which are as (6)-(9):

$$u = \sum x, w/2 \quad (6)$$

$$v = \sum y, h/2 \quad (7)$$

$$s = w * h \quad (8)$$

$$r = w/h \quad (9)$$

Here it has to be noted that these equations are non-linear, so the conversion is performed as pre-processing stage in the proposed method. Here the KF measurement function is also used as explicit function to handle the non-linearity.

Here in this computational approach the strategy considers pedestrian tracking dataset and further generates a matrix form of storage correspond to detection records  $d_{rec}[i]$ . Here the matrix form of  $d_{rec}[i]$  are generated using YOLOv4 object detector which is already defined in subsection 2.2. Here  $i$  refers to the detection indexes stored in  $d_{rec}[i]$ . Here the proposed strategy further performs conversion of  $d_{rec}[i]$  considering the  $Z \leftarrow d_{rec}[i]$  of bounding box coordinates to  $[u, v, r, h]$ . In this operation, the system incorporates an explicit function of bounding box detections  $F_{BBD}(\cdot)$ : pass in:  $d_{rec}[i], R$ . Here  $R$  refers to the covariance estimation for standard deviation of 1. The operation of bounding box conversion considering YOLOv4 object detections can be represented as (10).

$$[u, v, r, h] \leftarrow F_{BBD}(d_{rec}[i], R) \quad (10)$$

The state transition principle in the proposed strategy is built on the top of constant velocity model. However, the proposed study also talks about track handling schema considering linear constant-velocity KF for bounding box tracking. The idea of KF is used in this proposed approach in such a way where it tracks an

object mostly pedestrian entity using a sequence of detections or measurements to estimate the state of the object based on the motion model associated with that object. In a motion model state represents collection of various quantities by which status of an object can be represented. Here the state could be its position, velocity or acceleration. The state transition from times  $t_k$  to  $t_{k+1} = t_k + \alpha_t$  considers a constant velocity model such as (11):

$$X_{k+1} = \delta X_k \text{ where } \alpha_t \in \delta \quad (11)$$

However, in the existing systems the constant velocity model lacks effectiveness in describing the actual motion of the object of interest (i.e., pedestrians) in video. Also, it cannot properly handle the variations among aspect ratio and area state measures. However, the proposed approach of MOT handles these constraints by constructing an enhanced function of  $F_{TrackingKF}(\cdot)$ . Here the customized function initially enables the constant velocity Kalman filter for bounding box tracking and also measures the detections in a specific format. Further it initializes a null velocity and also estimates the velocity states by means of covariance measures. However, the system also models capturing unknown acceleration through which it constricts the KF for tracking considering both motion model and state transition model along with measurement model.

The proposed strategy also further evaluates the measurement association ( $a_T$ ) for each track ( $T_i$ ). Here the strategy associates the track with appropriate measurements. Here each detection should be associated with the existing tracks which are maintained by the multi-object tracker. The tracking model basically hypothesizes an object of interest with a specific track value in the current frame. Further the system is using estimates its bounding box. All the detections and its associative tracks are further loaded in a file structure called as ( $A_{MAT}$ ). It has to be also noted that during Kalman Filter prediction, the counter of track value is incremented and once the track is associated with the measurement then it is reset to 0. Here the system also considers estimation of predefined maximum age of a track ( $A_{max}$ ). Here the system is modeled in such a way where it sees if tracks have exceeded the predefined limit of  $A_{max}$  or not if it exceeds then it considers that the tracks might have left the scene and deleted from the track set. This is how the system also ensures memory utilization efficiency during track management. Further the KF model is initiated for new track hypotheses which correspond to multiple pedestrian detection which are not associated with the existing tracks.

## 2.5. Formulation of allocations metrics

The study in this phase of research formulates allocations problem where it applies the policy of Hungarian algorithm to build association between predicted Kalman states and newly arrived measures. The proposed strategy basically utilizes three different types of distances for allocations strategy which are: bounding box intersection over union (BBIOU), Mahalanobis distance measure, and appearance cosine distance. Here in the proposed approach of MOT, the distance of BBIOU refers to the measure between a track and detection based on the overlap ratio of two bounding boxes which could be defined as (12).

$$d_{IoU} = 1 - \frac{I_{AREA}}{U_{UNION}} \quad (12)$$

On the other hand, the measure of Mahalanobis distance is another evaluation criteria through which the distance between detection and tracks are evaluated considering a statistical evaluation of probability density functions. However, the evaluation of distance Mahalanobis can be estimated considering the following expression.

$$d_{Mahalanobis} = (B - Hx)^T S^{-1} (B - Hx) \quad (13)$$

Here  $B$  represents the measures associated with bounding box,  $x$  represents the track state and  $H$  refers to Jacobian measurement functionality for dimensional space measure. On the other hand,  $Hx$  refers to the predicted measurement. Here  $S$  denotes covariance matrix. However, the study also evaluates the distance between the detection and predicted tracks considering appearance feature vector which is also referred as appearance cosine distance.

## 2.6. Formulation of cost metric

The proposed idea in this research approach basically combines both  $d_{Mahalanobis}$  and appearance feature cosine distance to estimate the set of new detections to the set of current line of movement (LoM). The cost estimation in proposed study is a minimization problem which is formulated as (14):

$$T_c = \sum \mu \times C_{Mahalanobis}, (1 - \mu) \times C(A_{cosine}) \quad (14)$$

where  $0 < \mu < 1$ .

Proper track management could also play a crucial role in the proposed study to handle the cost metric. This approach of cost evaluation with effective track management basically handles the occlusion in the proposed tracking model. In this regard, the study introduces a matching evaluation solution that prioritizes the more frequently observed objects over a video scene to assess the association likelihood.

### 2.7. YOLOv4 and person re-identification network modeling

The study in the proposed approach of MOT basically considers a dataset of pedestrian tracking where the detections are generated from the YOLOv4 object detector model. Here the system considers both ground truth data and the detections from the pre-trained YOLOv4. However, the proposed approach also further performs conversion of detections to bounding box coordinates using (11). Further the system also integrates a person re-identification network which is also a pre-trained network to evaluate the appearance feature for each detection. It has to be noted that for extraction of appeared feature vector correspond to each detection, the system considers bounding box coordinates and convert them to crop frame. Finally, the appearance vector of ( $A_{vec}$ ) is constructed considering a prediction strategy for person re-identification. Here the cropped frame object of interest image is used to predict the appearance features and the final  $A_{vec}$  is generated. Further the system of MOT considers these bounding boxes and  $A_{vec}$  and iterates them to evaluate the algorithm 1 and generates allocations which are further managed under LoM initiation, deletion and appearance feature update. The tracker also evaluates KF to predict tracks which are further utilized for generating allocations. Finally with these optimized iteration steps the system performs tracking of pedestrian movements. Further outcome of the proposed tracking model is evaluated considering different performance parameters which are multi-object tracking accuracy (MOTA) and multi-object tracking precision (MOTP).

#### Algorithm 1. Strategy for cost-effective allocation of tracks and track management scheme

Input: Track Indices( $T_i$ ), Detections ( $d_{rec}[i]$ ),  $A_{max}$   
Output: Track Management for minimum cost evaluation  
Begin  
1. Video Sequence Exploration Frame Reading  
2. Apply YOLOv4, Person Re-Identification  
3. Initialize unallocated detections ( $d$ )  
4. Compute:  $C_{Mahalanobis}$ ,  $C(A_{cosine})$  for  $T_k$  and  $d$   
5. Evaluate cost using (14)  
6. Apply Linear Allocations of cost  
7. Update all Allocations and unallocated tracks  
8. Update unallocated detection  
9. Initialize Predicted  $T_i$   
10. Formulate  $T_k$  for last ( $i-1$ ) frame instances  
11. For ( $i = 1$ )  
    a. Evaluate (Line-4 to Line-5)  
    b. For ( $i = i + 1$ )  
        i. Compute  $IoU$  cost for remining unallocated detections and tracks  
        ii. Linear Allocations  
    c. If ( $(i \leq A_{max})$ )  
    d. Evaluate (Line-11, Line-11a)  
12. Update cumulative Allocations with unallocated detections and tracks  
13. End of For  
End

## 3. RESULTS AND DISCUSSION

The proposed study considers evaluation of the MOT framework to justify the propositions applicability in real-time multi object tracking. Here the proposed tracking workflow model considers pedestrian tracking video data for the purpose of evaluation. The study further also considers object-oriented approach to evaluate the MOT tracker framework. Here the work-flow for the evaluation of the proposed MOT tracker model initially constructs a LoM log and further initializes a minimum score for detection of object of interest. Further it enables a processing looping structure to significant score of detections which are further used to run the Person re-identification network model. Here the tracker considers high-score detections to predict the tracking of LoM for the detected pedestrian movement on the street.

The study also maintains a record for the log of LoM which is further also used for evaluation of the performance metric. However, the evaluation measures further show how the tracking outcome in the proposed study model is comparatively better than the existing tracking models. In this regard the performance evaluation scores are computed in the measure of MOTA and MOTP along with false positive

(FP), false negative (FN), and recall scores. The study shows that the proposed workflow model for the MOT tracking accomplishes considerable MOTA score which is approximately 85.8% whereas MOTP score is found to be 94.5%. It also shows that the false positive score of 30 along with FN value of 62. However, the recall performance is found to be 90% which are comparable with the existing baselines. The comparative outcome as highlighted in the Figure 1 clearly shows that the proposed tracking model outperforms the baseline approaches of k-dense neighbors based tracker (KDNT) [28], simple online and real-time tracking (SORT) [29] and lifted multicut problem (LMP) [30] in the measure of both MOTA and MOTP. The prime reason behind this is the inclusion of cost-efficient track Allocations and management schemes. The comparative outcome is also visualized through Figures 1(a) and 1(b).

The traditional popular MOT methods by [28]–[30] also talk about the scope of multiple hypothesis (MHT) tracking along with probabilistic data association problem. Even though these approaches are found to cater the requirements for tracking accuracy when integrated with pruning schemes, but the efficient tracking-by-detection comes with increased computational and design complexity for association metric evaluation under frequent occlusions. Unlike existing baseline strategies of MOT, the proposed model of MOT considering the cost-effective approach for optimized trajectory formulation addresses this challenging scenario in conventional approaches and improvise that to retain well balance between both complexity and MOTA. The study model also claims its effectiveness in implementing simplified work-flow design which makes the operations more cost-effective and resource friendly when compared with the baseline studies of MOT. The comparative analysis also shows that the efficient track management scheme in the proposed system not only improves its operational and design efficiency from the cost of computation point of view but also ensures higher reliability in ensuring better multiple target tracking accuracy and precision which makes it more suitable to be implemented over real-time scenarios.

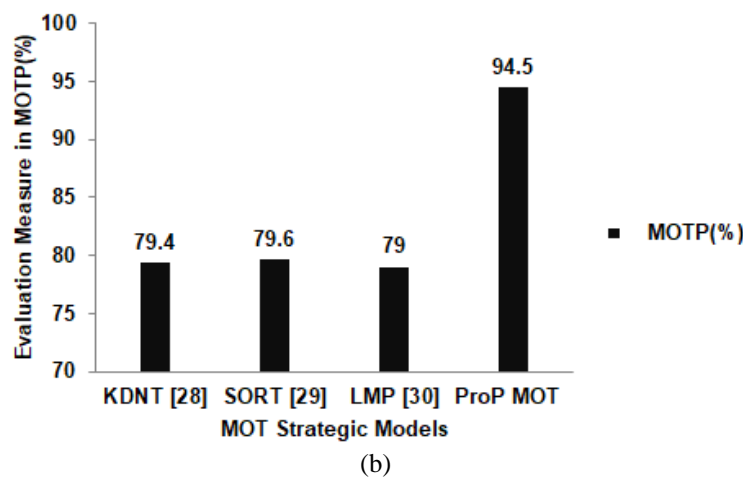
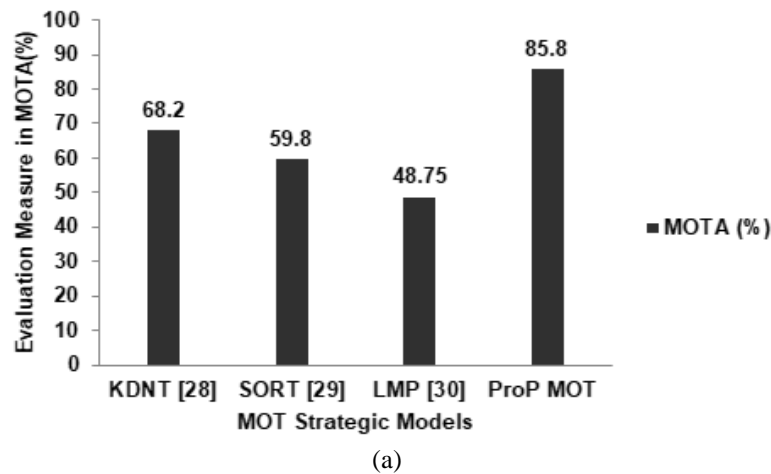


Figure 1. Comparative analysis (a) MOTA (%) and (b) MOTP (%)



#### 4. CONCLUSION

The proposed study introduces a cost-effective MOT framework where it integrates both YOLOv4 and Person re-identification network with a multi-object tracker design considering the appearance features to improve the performance of pedestrian tracking. Here the system formulation also evaluates problem of allocations distances where implements a unique approach of allocations strategy with efficient track management scheme where it also evaluates the cost considering both Mahalanobis cost and cosine cost. Here the proposed object tracking algorithm design is modeled as simple as possible from the design point of view and also person re-identification entification deep learning model here contributes towards generating deep appearance features which improve the allocations handling schema with track management. Here integration of appearance information into data association also handles the problem of occlusion to a higher extent over longer period of time. The outcome of the study shows that the proposed model outperforms the existing baseline tracking strategies with 85.8% MOTA and 94.5% MOTP scores making it suitable for real-time tracking of object of interest. The scope of the study lies in modifying the Mahalanobis distance in future so that the distribution of covariance can be handled more appropriately in the presence of occlusion.




#### 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] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, "Deep learning in multi-object detection and tracking: state of the art," *Applied Intelligence*, vol. 51, no. 9, pp. 6400–6429, Sep. 2021, doi: 10.1007/s10489-021-02293-7.
- [3] Bo Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 1918–1925, doi: 10.1109/CVPR.2012.6247892.
- [4] S. K. Pal, D. Bhounik, and D. Bhunia Chakraborty, "Granulated deep learning and Z-numbers in motion detection and object recognition," *Neural Computing and Applications*, vol. 32, no. 21, pp. 16533–16548, Nov. 2020, doi: 10.1007/s00521-019-04200-1.
- [5] D. B. Chakraborty and S. K. Pal, *Granular video computing with rough sets, deep learning and in IoT*. World Scientific, 2021.
- [6] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, Apr. 2012, doi: 10.1109/TPAMI.2011.155.
- [7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 3354–3361, doi: 10.1109/CVPR.2012.6248074.
- [8] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
- [9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4.
- [10] T.-Y. Lin *et al.*, "Microsoft COCO: common objects in context," In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, vol. 8693. Springer, Cham. 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1\_48.
- [11] A. Kuznetsova *et al.*, "The open images dataset V4," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, Jul. 2020, doi: 10.1007/s11263-020-01316-z.
- [12] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.
- [14] J. Hussain, B. R. Prathap, and A. Sharma, "An improved and efficient YOLOv4 method for object detection in video streaming," In: Shukla, S., Gao, X.Z., Kureethara, J.V., Mishra, D. (eds), *Data Science and Security. Lecture Notes in Networks and Systems*, vol. 462. Springer, Singapore, 2022, pp. 305–316, doi: 10.1007/978-981-19-2211-4\_27.
- [15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
- [16] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.
- [17] L. A. Zadeh, "A note on Z-numbers," *Information Sciences*, vol. 181, no. 14, pp. 2923–2932, Jul. 2011, doi: 10.1016/j.ins.2011.02.022.
- [18] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: scaling cross stage partial network," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 13024–13033, doi: 10.1109/CVPR46437.2021.01283.
- [19] K.-C. Liu, Y.-T. Shen, and L.-G. Chen, "Simple online and realtime tracking with spherical panoramic camera," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2018, pp. 1–6, doi: 10.1109/ICCE.2018.8326132.
- [20] S. Menshov, Y. Wang, A. Zhdanov, E. Varlamov, and D. Zhdanov, "Simple online and realtime tracking people with new 'soft-iou' metric," in *AOPC 2019: AI in Optics and Photonics*, Dec. 2019, doi: 10.1117/12.2547922.
- [21] O. Urbann, O. Bredtmann, M. Otten, J. P. Richter, T. Bauer, and D. Zibiczky, "Online and real-time tracking in a surveillance scenario," *arXiv preprint arXiv:2106.01153*, 2021.
- [22] X. Hou, Y. Wang, and L.-P. Chau, "Vehicle tracking using deep SORT with low confidence track filtering," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Sep. 2019, pp. 1–6, doi: 10.1109/AVSS.2019.8909903.
- [23] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *International conference on machine learning*, pp. 448–456, 2015.
- [24] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.




- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.
- [26] H. Yu, X. Li, Y. Feng, and S. Han, "Multiple attentional path aggregation network for marine object detection," *Applied Intelligence*, vol. 53, no. 2, pp. 2434–2451, Jan. 2023, doi: 10.1007/s10489-022-03622-0.
- [27] P. R. Gunjal, B. R. Gunjal, H. A. Shinde, S. M. Vanam, and S. S. Aher, "Moving object tracking using Kalman filter," in *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, Feb. 2018, pp. 544–547, doi: 10.1109/ICACCT.2018.8529402.
- [28] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "POI: multiple object tracking with high performance detection and appearance feature," In: Hua, G., Jégou, H. (eds), *Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science*, vol. 9914. Springer, Cham. 2016, pp. 36–42, doi: 10.1007/978-3-319-48881-3\_3
- [29] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3464–3468, doi: 10.1109/ICIP.2016.7533003.
- [30] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3701–3710, doi: 10.1109/CVPR.2017.394.

## BIOGRAPHIES OF AUTHORS



**Divyaprabha**    is associate professor at Sri Siddhartha Institute of Technology, Tumkur, India. She received her bachelor's degree from Bangalore University and master's degree from BITS, Pilani, Rajasthan. She is a member of ISTE. Her field of interests are image processing, machine learning. She has published papers in conferences and journals. She can be contacted at email: divyasy74@gmail.com.



**Guruprasad Seebaiah**    is associate professor at Sri Siddhartha Institute of Technology, Tumkur, India. He received his bachelor's, master's and doctoral degree from Visvesvaraya Technological University (VTU), Belgaum, India. He is a member of ISTE. His field of interests are image processing and biomedical instrumentation. He has published research papers in reputed conferences and international journals. He can be contacted at email: guruprasads@ssit.edu.in.