




Modified YOLO Module for Efficient Object Tracking

Varsha Kshirsagar Deshpande , Raghavendra Bhalerao  and Manish Chaturvedi , IITRAM, Ahmedabad

Abstract—In the proposed work, initially, the YOLO algorithm is used to extract and classify objects in a frame. In the sequence of frames, due to various reasons the confidence measure suddenly drops. This changes the class of an object in consecutive frames which affects the object tracking and counting process severely. To overcome this limitation of the YOLO algorithm, it is modified to enable and track the same object efficiently in the sequence of frames. This will in turn increase object tracking and counting accuracy. In the proposed work drastic change in confidence scores and class change of an object in consecutive frames are identified by tracking the confidence of a particular object in the sequence of frames. These outliers are detected and removed using the RANSAC algorithm. After the removal of the outliers, interpolation is applied to get the new confidence score at that point. By applying the proposed method a smooth confidence measure variation is obtained across the frames. Using this, average counting accuracy has been increased from 66% to 87% and overall average object classification accuracy is in the range of 94 - 96% for various standard datasets.

Index Terms—Object detection, YOLO, Motion Tracking, RANSAC.

I. INTRODUCTION

The object tracking system is an interesting field for researchers in different applications like robot navigation, unmanned aerial vehicle [1], driver assistance with vehicle tracking module [2], hand gesture recognition [3] in human-computer interaction and music transcription using finger tracking [4]. Recent studies reflect many complex problems effectively handled by deep learning techniques such as organoid in vitro model detection and tracking for the reduction of labor cost and precise surveillance [5]. The advancement of deep learning brings exclusive improvement in object detection and improves tracking approach, which includes object extraction and tracking of the extracted target. Correct object detection is at high peaks in many applications. Change detection of a particular target through video frames using deep learning developments has improved a lot in video detection technology. To link the same object throughout the frames, a data association algorithm based on spatial information or appearance is enforced here. Output from object detection is used as input credentials for the tracker [6]. A part-based particle filter is implemented using the hidden state of center of the vehicle for vehicle tracking. Here a pre-trained geometric model with rich information from invalid parts makes precise predictions. From motion patterns, accuracy of tracking vehicle is improved for

this algorithm. Vehicles are assumed to have similar fixed sizes and the ground is assumed to be flat [7], which is practically not possible since the area of the vehicle depends on the position of the camera [8]. In the literature, many techniques such as the nearest neighbour, multiple hypothesis tracking, joint probabilistic method are available, where data is available for different applications, but the drawbacks of the above-mentioned algorithms remain as it is. These drawbacks include, lack of robustness for the nearest neighbour, the requirement of prior data increases complications for data association in joint probabilistic algorithm and multiple hypothesis implementation require large computations [9]. As per the author, in traditional methods like template matching or kernel-based methods, object detection gives the initial position of the object and tracking can be considered a process of detecting a target based on initial information. In the long range application, it increases computation and thus degrades the output due to changes caused by shadows and illumination effects [10]. Careful and smooth use of big data-set, as well as the powerful Graphics Processing Unit with quality results increases the importance of deep learning in the field of computer vision applications like object detection and classification. The backbone of deep learning is a convolutional neural network that has an automatic feature extraction which gives quality results [11]. So from the deep learning colony YOLO module is preferred for object detection and classification. The N-YOLO module divides the image instead of resizing it and the correlation-based tracking algorithm is used after merging. The computation time for detection and tracking can be decreased here. But the limitation of the above model is that when the tracker gets the object back, new ID gets assigned and the algorithm is not able to handle the re-entering of the object [12]. YOLOv3 is a modified version of the YOLO algorithm which can detect and classify multiple objects with a single inference and hence the computation time required for the same is less and also the accuracy of the algorithm is increases as compared to YOLO [13].

Specifically in Vehicle tracking for vehicle counting application, YOLO detects target accurately, however it is a challenging task since 1. The area of the vehicles varies with the position of camera. 2. Sometimes, the color of the vehicle and color of the road matches and fewer features are available due to gray level variation. 3. While entering or leaving the video, a little part of the vehicle appears in the video which can affect the result of classification. 4. If a heavy vehicle covers a small vehicle, the class of the vehicle changes. 5. Occlusion and surface variation can affect the output in terms of false-positive value 6. Morphological variation of the

Varsha Kshirsagar Deshpande, Ph.D Scholar, Raghavendra Bhalerao, Assistant Professor, Manish Chaturvedi, Assistant Professor, Institute of Infrastructure Technology Research and Management Ahmedabad e-mail: varsha.kshirsagar.19pe@iitram.ac.in

vehicles between two frames is significant. Due to all these reasons the confidence score of the detected object reduces below the threshold and the class of the object changes for such frames which can affect accuracy of the object counting. The vehicles at junctions can take different directions and need to be tracked along the entire roundabout to find the entry and exit. Generally an object starts tracking at entry point and is tracked until exit point and once it crosses a specific entry and exit it is counted in that direction. In such case if the confidence score is not identical at entry and exit point it will affect the counting accuracy. Based on this, the primary research objective of the proposed work is to design an efficient classification tracking module for the extracted target with an acceptable confidence score in the video streams. The initial stage of the work focused on object recognition and classification, which are taken care by YOLO module. Influence of variation in confidence score, area, velocity due to occlusion and shadow is pulled off by rejecting outliers using the RANSAC algorithm and these rejected points are updated by curve fitting using linear interpolation.

Our main contributions in the proposed work are; 1. Tracing target object path by importing features such as confidence score. 2. Incorporate tracing by improving the low confidence score of the extracted classified target with smooth curve fitting using the RANSAC algorithm and linear interpolation. 3. To improve counting score by rectifying misclassified vehicles. In the proposed work section 1 and 2 summarises the introduction and related work of the proposed module. Section 3 contains the working flow of the algorithm under the heading of methodology. Section 4 consists of experiment and result discussion. Section 5 comprises Conclusion.

II. LITERATURE SURVEY

Literature survey reviews development of object detection and tracking algorithm in last decades. A systematic survey of counting objects from digital images is discussed by dividing available techniques and main tools into eight sections [22]. Object detection techniques can be broadly divided into traditional approach and deep learning approach. In traditional approach, appearance based method detects object based on color or depth. Whereas in motion based approach, background subtraction, spatio temporal filtering and optical flow are the strategies used to extract object from image. For extraction and automatic classification of scene images, modified binary local and global descriptor is used for effective results [23]. Segmentation of similar regions can be utilized for object extraction with mean shift clustering and camshift algorithm for object tracking which is recognized for good processing efficiency and its simplicity [24]. Highlighted background subtraction methods from the literature are, frame difference, region based, texture based background frame, gaussian mixture based and markov model. These can be upgraded for adaptive nature to update background frame after certain interval. In optical flow technique image optical flow field is calculated and clustering is performed as per optical flow distribution of the image [25]. Introduction of deep learning in the era of the object detection is as a regression module [26], afterwards it is modified,

by replacing last layer of alexnet with regression layer for the object detection as well localization. With deep multiBox method multiple objects are detected with its localization [27]. In the next version sliding window approach with multi-scaling approach is used for detection, classification and localization [28]. In the upgraded version the image is divided into small regions and the application of deep CNN gives feature vector where support vector machine is used for classification [29]. "You Only Look Once (YOLO) is popular and now frequently used module for object detection due to its features [13]. YOLOv3 gives a trade-off between accuracy and speed [16]. In the field of object detection still results of small object detection are not satisfactory by learning shallow features at the shallow level and deep features at the deep level, the proposed Multi-Scale feature YOLO learning (MSFYOLO) tool is used for better results [30]. YOLOv4 is modified with denseNet framework which is used to study echocardiographic images for diagnosing congenital heart diseases (CHDs) [19]. Tracking is to establish association between the same target through the video by crucial parameters. Initially an association between same target is maintained by a point correspondence and the method is comfortable for the small objects, where points need to be tracked in every frame. The deterministic and probabilistic are considered as two types of point tracking. As per literature here, calculation of the missing point is handled by hypothetical point [31]. In another method, the background subtraction is used for target extraction, then centroid trajectory is utilized for establishment of relation between target for tracking [32]. The Kernel Tracking is the another way of data association which is based on the template matching or appearance of the target and can be used for single tracking or the multi-tracking. The kernel correlation filter utilizes number of training samples and improve quality of tracker. This method is classified into parametric and non parametric. In the parametric distribution at each frame, the target location is upgraded through statistical approach such as a mixture of gaussians or its upgraded version with variable parameters and spatial mixture method. The color identification in the Hue Saturation Value (HSV) color space and active contour models with open source software is used for association and recognition of object in spite of variation in size and shape [33]. Expectation maximization is another mathematical tool used in the literature. In the extended version, methods like poisson distribution, dirichlet distribution and the regression models are also considered as statistical model for capturing motion parameters. But to reduce the high computation cost and complexity between the object module with hypothesized position for a non parametric model, the mean shift algorithm is used and the module is upgraded with a weighted histogram. In the modified module a spatial- color histogram is used, but here it is required that some part of the object should be inside the selected shape whose location is defined by previous position of the object. To wipe out such requirement, the kalman filter or the particle filter can be used to predict the location of the target in next frame [34].

The silhouette tracking is observed when the target is available in the form of complex shape. Here the data association is obtained through a model, developed using previous frame

TABLE I
SUMMARY OF VARIOUS METHODS

Sr No.	YOLO-version	Application	Remark
1	YOLO	A Modified YOLO Model for On-Road Vehicle Detection in Varying Weather Conditions	The ML-based classifier is used to classify Vehicles and 16 convolutional layers of YOLO has been used along with two fully connected layers at the end [14].
2	YOLOv2	Vehicle Logo Detection Based on Modified YOLOv2	Here clustering of the bounding box of the vehicle logo database, reconstructing network pre-training and multi-scale detection training are used in a modified version [15].
3	YOLOv3	An Accurate and Fast Object Detector Using Localization using Gaussian YOLOv3	Improvement in accuracy is obtained through Gaussian modeling, loss function reconstruction and the utilization of localization in YOLOv3network [16].
4	YOLOv3	Robust Thermal-Visible Heterogeneous for face recognition	Here YOLOv3 provides an advanced solution in face recognition for thermal and visible imagery for security purposes also modified CycleGYAN module is used to translate LWIR images to visible images with good robustness and efficiency [17].
5	YOLOv4	Face Recognition Approach Based on a Cycle	YOLOv4 module along with the image super-resolution module (ISR) and AI method is used to detect the wearing of a helmet [18].
6	YOLOv4	Generative Adversarial Network	A Deep Learning module was introduced for echocardiographic image detection of VSD using YOLOv4 DenseNet framework [19].
7	YOLOv4	Using YOLOv4 version accuracy using Coverage Ratio of Street Trees improved.	Here YOLOv4 is used for object detection on street trees where the coverage ratio can be estimated using parameters. Integration of remote sensing images is utilized to improve the accuracy of the coverage ratio [20].
8	YOLOv5	Application of an Improved YOLOv5 Algorithm in Real-Time Detection of Foreign Objects by Ground Penetrating Radar.	The small object detection problem of YOLOv5 is improved by the modified network structure [21].

parameters and constructed shape model for the consecutive frames using similar shape description. This shape model may be in the form of a line, an edge, a color histogram or object contour. According to the model it is divided into two approach, shape matching and contour. Silhouette tracking can handle variety of shapes and the binary indication represents the object as one and non-object part by zero which makes the system comfortable for further processing. Occlusion handling and capability of dealing object with split and merge is quite complex [35]. Introduction of the motion concept for association, with coherent tracking can solve issue of occlusion but here number of objects are assumed to be the same. For a multi-frame approach to preserve temporal coherence of speed and position, first detection of object is necessary for good result [36]. The real world multiple object detection system is introduced with high processing speed and accuracy [37].

YOLO's initial version has two limitations, inaccurate positioning and the other one is lower recall rate. YOLOv2 improves these parameter and comes out to be a better and faster version. YOLOv3 includes multi-scale features for object detection and adjusts the basic network structure. YOLO V4 gives focus on comparing data including previous version features with improved performance. Multiple network architectures of YOLO V5 are more flexible to use, have a very lightweight model size and are on par with the YOLO V4 benchmark in terms of accuracy and more suitable for long distance real time detection. However, YOLO V5 is popular because it is less innovative than YOLO V4 and have flexible control of model size hence reflects with performance improvement among the different versions of YOLO [38]. YOLOv6 gives better trade-off in terms of accuracy, speed and better mean average precision for real time applications preferably in industrial applications [39]. However, all the YOLO versions have a limitation that, due to various reasons the confidence score changes drastically for few of the frames and which changes the class. Table I represents various modified versions of YOLO family for various applications. In the proposed method, the confidence score with a classified deep

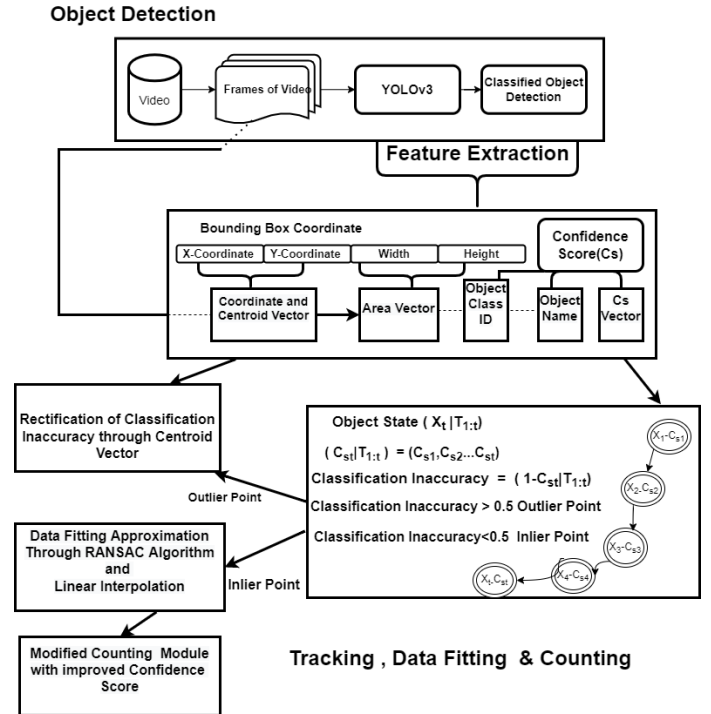


Fig. 1. System Block Diagram.

learning module is precisely utilized for data association and tracking of the object through the video. The confidence score will add strength to the existing parameters for data association and following the same target, also it will rectify classification inaccuracy and can be extended for object counting.

III. METHODOLOGY

Following are the main steps to implement the algorithm. Figure 1 represents System Block Diagram. The steps are elaborated in following section.

- 1) Object detection using YOLO
- 2) Feature Extraction



Fig. 2. Yolo Output.



Fig. 3. Extracted Vehicle in consecutive frames with class change and drastic change in Confidence Score.

- 3) Outlier rejection
- 4) Data fitting
- 5) Object tracking / counting

A. YOLO Object Detection

Accurate object detection and trailing the same detected object throughout the video is an important aspect of motion tracking. The introduction of deep learning through the YOLO module provides a reliable solution in the field of object detection. YOLO works well for individual frames as shown in Figure 2 Tracking through feature extraction is one of the reliable methods to follow the trajectory of the same object through the video. while tracking an object in a video using a feature such as confidence score, it may result in class change or the object may not get detected in that particular frame due to variation in its confidence score. This may be the outcome of various reasons across the frames thus making it difficult to track the object in consecutive frames. Another limitation of YOLO is that it does not provide a unique object ID to the object which makes it further difficult to track the object in a video. In the proposed approach we are using the information from consecutive frames based on the confidence score and object position. For example, due to some reasons a truck can be misidentified as car in next frame, and after a few frames it can be identified as truck again. This ambiguity makes it difficult to count and track objects in a video sequence. In the Figure 3, the highlighted portion shows the extracted car from consecutive frames. On implementing YOLO classification algorithm, the object is correctly identified as a car till third frame but in fourth it is reflected as a bus and again classified as a car in the next frames. This classification inaccuracy may be a result of shadow effect. To overcome this limitation, in the proposed methodology we used the YOLO algorithm on a frame and observed its variation over the sequence of the video. YOLO does not consider the neighbouring frames even though these are highly correlated. The confidence measure of consecutive frames can be used for overcoming this limitation. Features extracted from the YOLO algorithm are height, width and centroid of the bounding box with classified output giving confidence Score . Multiple features from YOLO algorithm increases the robustness of tracking the same object through the video frames. Implementation of the YOLO algorithm on video database gives classified

output with object ID and confidence score. While tracing a particular object through the video, for some cases variation in confidence score occurs below the threshold due to certain reasons such as occlusion and background variation. This results in classification inaccuracy and the change of class is reflected in the output. Hence tracking can be divided into two categories, tracking with error and tracking without error. The proposed algorithm uses extracted parameters to overcome this classification inaccuracy. The extracted parameters along with the centroid smooth curve of the area is utilized to track particular targets throughout the video.

B. Feature Extraction

To follow the same detected object through the frames of the video, extracted features are considered one of the most convenient and reliable methods of tracking. Multiple features are a crucial contribution from the YOLO algorithm which gives us features like, class and centroid of the extracted object along with width and height of the bounding box. Object name and object class ID with its detection confidence score are also considered as exclusive parameters for tracking the extracted target throughout the frames of the video. The area of a bounding box is calculated with the help of the width and height of the bounding box which are considered to be supporting parameters to follow the trajectory of the detected target.

C. Outlier Rejection

YOLO shows excellent accuracy while extracting and classifying objects from image. While tracking a target through the frames of a video, due to variation in parameters or shadow/illumination effect the confidence score of extracted target for that particular frame drops suddenly which results in changes in the class as shown in Figure 4. This represents variation of confidence score and class ID with respect to frame number, while tracking a car through the frames, near frame number 410 and 430 suddenly confidence score moves to zero which indicates object is missing in that particular frame. Also, for some frames after 510, as represented in blue color in the figure, the class of the object changes with decrease in the value of confidence score. This in turn, changes the class of object from car to bus. Variation in confidence score detects particular frames for which class of the object get changed and missing class for missing data is as shown in the figure. YOLO gives the confidence score for each detected object. Higher the confidence score better is the accuracy. The maximum value of the confidence score is 1 when an algorithm has 100 percent confidence about the accuracy of an object. However, this number is rarely met and generally a score is between 0.8 to 0.9 for good detection accuracy. Based on assumption, the relation between classification inaccuracy and confidence score is given by

$$1 = ConfidenceScore + ClassificationInaccuracy \quad (1)$$

YOLO gives confidence score to each detected object and classifies the object into that particular class with highest confidence score. Based on confidence score, detected object

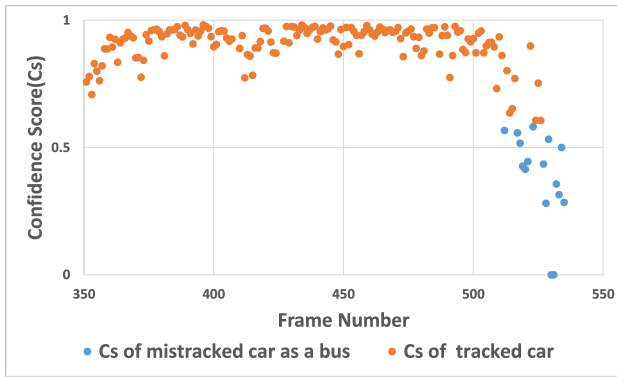


Fig. 4. Confidence score vs Frame Number.

can have more than one class, among which class having highest confidence score is considered. This information of confidence score is passed for highly accurate classification. Multiple extracted features from object detection algorithm in a proposed algorithm are deployed for reduction of classification inaccuracy. The trajectory of extracted target centroid (X-coordinate ,Y coordinate) of the target along with its confidence score gives smooth curve if the classification of extracted target is correct. But, variation in parameters from YOLO object detection algorithm drops the confidence score below threshold for that particular frame which is normally considered as 0.5. Due to which change in class takes place and classification inaccuracy increases as per above mentioned equation 1. Abrupt variation reflects in graph as shown in Figure 4. As per the extracted features and observation during above mentioned variation in confidence score(C_s), other than C_s no transient variation is observed in remaining parameters. These parameters are utilized to rectify classification score for restoring class ID back. While extracting target object through the frames when $d = C_s(n + 1) - C_s(n)$ where, d is the difference between confidence scores of the same target of consecutive frames. If the value of d is high there tends to be a change in the class. Algorithm to retain class of extracted target by YOLO if the confidence score is less than threshold is as described in Figure 5 Initially, to track the object throughout the video, it is converted into frames. Number of objects from previous frame p_k and current frame q_k are compared by considering class of the object, where n and $n - 1$ are the current and previous frames of the video and k is the number of objects from a frame. If the class of both the objects is same then distance between their centroid is compared with minimum distance condition and same object in next frame is replaced. Here another parameter such as area of the same object for previous frame and current frame is also considered. Area of object varies when it moves away or towards the position of the camera. As per experimental work, area of the object in current frame is between 80 to 90 % of the area of object in previous frame. This adds leverage to catch the same object in next frame. Once object is identified its confidence score and classification inaccuracy is calculated and if it is greater than threshold class id of the same object, it is saved and compared with previous one assigned with previous class id by removing class change problem. Figure 6 represents centroid tracking for missing or class change object

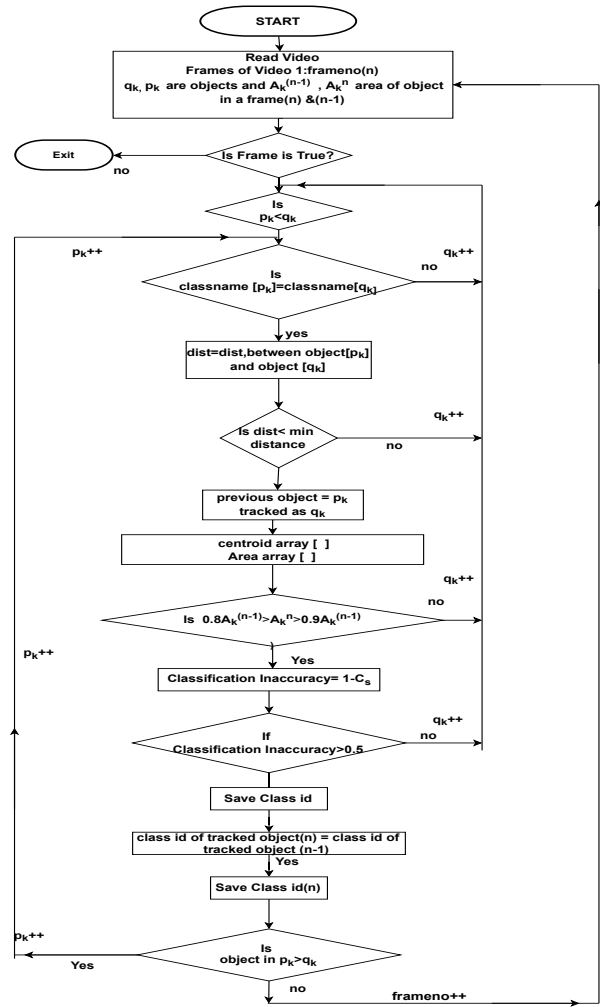


Fig. 5. Algorithm.

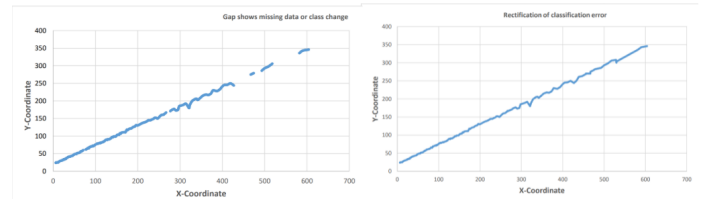


Fig. 6. Tracking through centroid (missing/class change and recovered).

which is recovered using proposed algorithm.

Confidence Score is represented by C_s and object state confidence score is represented by

$$(C_s^t \mid T = 1, 2 \cdot t)$$

t = time for which object will remain in video

Time Step = T

Vehicle State = X_t

Confidence Score of the object = C_s

Pr denotes probability of the particular

class and IOU indicates Intersection over Union

Confidence Score = $\text{Pr}(\text{object}) * \text{IOU}$

$0 < Pr(Object) < 1$
 $0 < IOU < 1$
 Maximum value of confidence Score =1
 $C_s = \text{Objectness Score} + \text{Classification Inaccuracy}$
 Error exist due to less features available due to occlusion,
 shadow, illumination .etc
if classification Inaccuracy > 0.5 **then**
 Class id(n) = class id(n-1)
 Append Confidence Score for tracking curve
 in RANSAC algorithm as a inlier point
 Else
 Outlier point
 Replace outlier points with 0
 Linear interpolation on output
 Plot
 Smooth Curve with improved confidence score
end if

As explained in the YOLO module, confidence score is the multiplication of probability of object and intersection over union.

The value of both the factors lies between 0 and 1. Hence the product of both will give a value ranging between 0 and 1. So the maximum value of Confidence score is 1. Threshold value can be varied to set objectness score, which can be defined as probability of presence or absence of an object in bounding box. In the given work if its value is greater than 0.5 then it is considered as inlier point otherwise it is considered as an outlier point as shown in Figure 7. For every frame of video in which extracted object travels, confidence score is appended in vector with the flow of video as in equation (4),

D. Track Approximation Through RANSAC Algorithm

RANSAC algorithm gives a solution to minimize effect of outlier points and gives the best fitted model with inlier points. The idea of the algorithm is to propose a model from the extracted data with variation in confidence score due to shadow, occlusion and many more parameters. RANSAC assists to search a perfect curve from extracted data of confidence score by YOLO. We can write, $\Gamma_t \subseteq R^n$ where Γ_t denotes tracking approximations with inlier points. Measurement Space is given by $\phi \subseteq R^n$ and confidence score countable data with respect to time is considered as $C_s \subseteq \phi$ for tracking approximations C_s with parameter vector. Measurement Space is given by $\phi \subseteq R^n$ and confidence score countable data with respect to time is considered as $C_s \subseteq \phi$. Tracking approximations are given by C_s with parameter vector. An outlier point is considered as datum that is distant from considered model affected by variation in parameters. Following equation distinguishes inliers and outliers.

$$d_{\theta t}(C_s) = \sqrt{(C_s - C_{\theta t})^T \cdot (C_s - C_{\theta t})} \quad (2)$$

Where $C_{\theta t}$ is the orthogonal projection of C_s onto the surface defined by model θ_t .

$$\theta_t = [C_s 1, C_s 2, \dots, C_s n]^T \subseteq R^n \quad (3)$$

Data preprocessing produces a data set as shown in equation (4), As shown in Figure 8, RANSAC algorithm will give

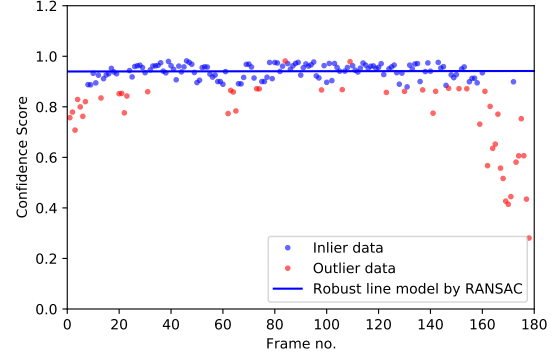


Fig. 7. Outlier rejection by RANSAC.

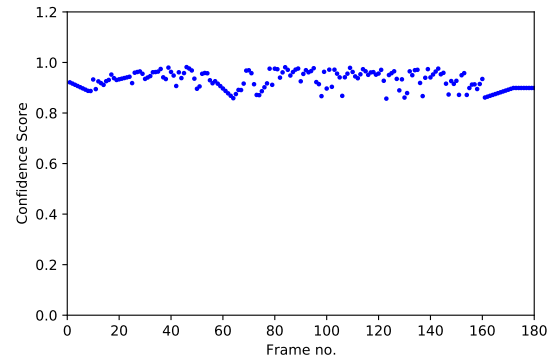


Fig. 8. Track approximation by RANSAC and Interpolation.

best appropriate curve by rejecting outlier points obtained by confidence score. From the graph, modified values can be updated for missing and mis-classified frame, Hence tracking curve can be improved with modified values.

$$\epsilon \begin{cases} > 0.5 & \text{Outlier point} \\ < 0.5 & \text{Inlier Point} \end{cases} \quad (4)$$

E. Linear Interpolation

Linear interpolation can identify the nearest point from the collected information of outlier and inlier points from RANSAC approximation. Initially outlier points are replaced by zero. Afterwards, linear interpolation is used to fill these zeroes by appropriate values. By using these values a model is built as shown in Figure 9 and 10. For all inlier points model is tested. The number of iterations are finalized when the algorithm gets sufficient number of inlier points to support the model.

F. Object Tracking/Counting

As mentioned in section II, YOLO provides excellent results for object detection but in case of classified counting, the output gets affected due to sudden change in class which is an impact of decrease in confidence score. After the vehicle is detected by YOLO, marker lines are inserted at entry or exit of a given video as shown in Figure 12 for the counting of

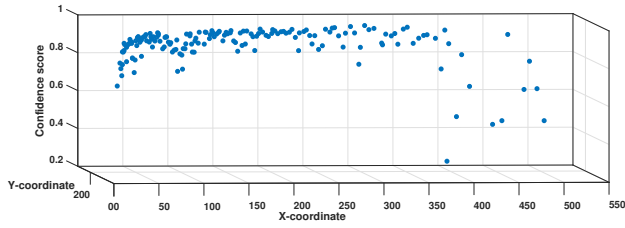


Fig. 9. Tracking through Confidence Score.

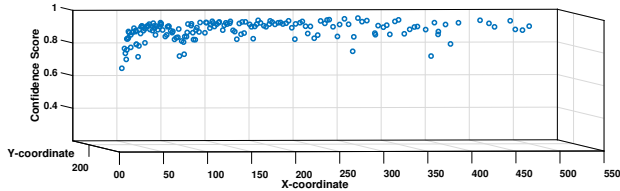


Fig. 10. Tracking through improved Confidence Score.

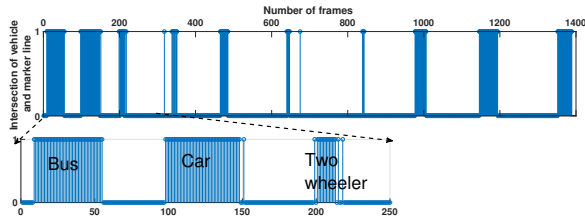


Fig. 11. Intersection Graph.

vehicles. Intersection of each vehicle with marker line initiates counting. Successive non zero outputs from the intersection of vehicle with marker line increments counter by one. Non zero outputs can be graphically represented in unique way as shown in Figure 11 to give exact count of the video. If the position of marker lines exists where class of the object changes or disappears due to decrease in confidence score as shown in Figure 14 then it will affect the accuracy of classified count. The intersection of two sets includes the points that are in all of the sets. Intersection is used to find the common points. In the given work, marker line is one of the set and other set is moving vehicles. If A and B represents vehicle and marker line, index vectors of these two vectors are i_a and i_b . Intersection will give a value which is common to both A and B, as well as the index vectors i_a and i_b as shown below,

$$\begin{aligned}
 C &= A(i_a) \\
 C &= B(i_b) \\
 [C, i_a, i_b] &= \text{intersection}(A, B) \\
 \text{FinalCount} &= 0 \\
 C &\leftarrow \text{count} \\
 \text{if } C \neq 0 \text{ then}
 \end{aligned}$$

count = 1

else

count = 0

end if

For every transition of count from 0 to 1

Final Count = Final Count + 1



Fig. 12. Frame with marker lines.



Fig. 13. Turning Movement in T-section data-set.

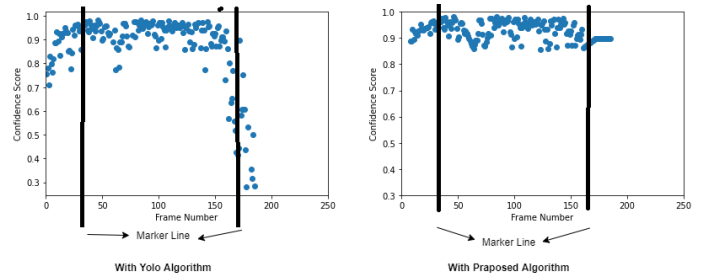


Fig. 14. Variation of confidence score, during counting with marker lines.

IV. RESULT AND DISCUSSION

Table II represents the counting of vehicles for our dataset with marker lines using conventional YOLO and proposed algorithm. The result of the count, itself indicates variation in confidence score which changes class affecting the counting accuracy drastically. The confusion matrix in Figure 15 and 16 reflects the implementation of proposed algorithm which rectifies missing object as well as class change variation and increases average counting accuracy from 68% to 88%. Figure 9 and 10 represent tracking through confidence score by YOLO module and with improved confidence score by proposed work.

The proposed algorithm can be effectively used in counting of vehicles in T-section data-set as shown in Figure 13

TABLE II
VEHICLE COUNTING

Vehicle class	YOLOv3 (Overall Accuracy=0.6862)				Modified YOLO (Overall Accuracy=0.8823)			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Bus	1	1	1	1	1	1	1	1
Others	0.7254	0.4615	1	0.6315	0.8820	0.6666	1	0.8
Two Wheeler	0.6862	1.0	0.4666	0.6363	0.8820	1	0.8	0.8888
Car	0.9607	0.75	1.0	0.8571	1	1	1	1

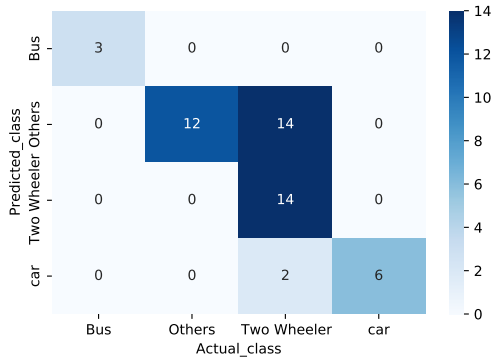


Fig. 15. Confusion Matrix by YOLO.

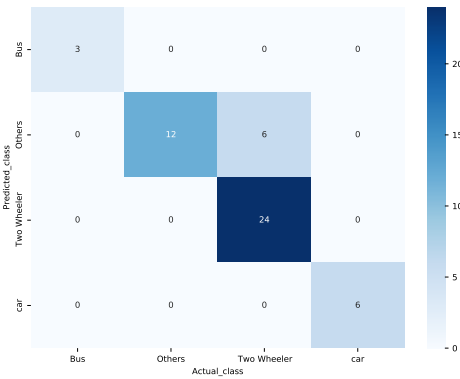


Fig. 16. Confusion Matrix by proposed algorithm.

where turning movement counting of vehicles severely affect accuracy if class changes during counting. The Proposed algorithm is implemented on three different datasets of Visual Tracker Benchmark. Figure 17 represents disappearance or class change of an object in successive frames of the video. Figure 18 represents the confusion matrix before and after proposed algorithm which also reflects how model accuracy and classification accuracy increases for different classes with variation in a scenario for own dataset.

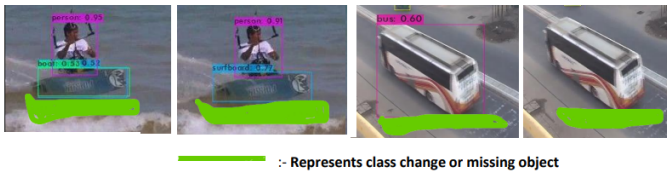


Fig. 17. Missing/class change objects in dataset.

Table III shows experimental analysis of various datasets along with our dataset. Precision, Recall, F-measure and Accuracy these are the four metrics used for analysis of proposed algorithm.

Implementation of the proposed algorithm reflects increments in above mentioned parameters for the extracted object throughout the video and increases object counting and classification accuracy from average 80% to 96.94%. From

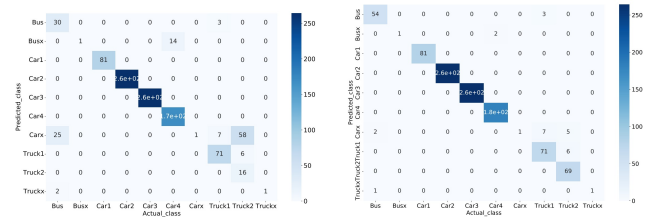


Fig. 18. Confusion matrix by YOLO and Proposed algorithm for OUR Dataset.

TABLE III
COMPARISON FOR VARIOUS DATASET

Dataset	Class	YOLOv3				Modified YOLO			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
HUMAN	Bus	1	1	1	1	1	1	1	1
	car1	1	1	1	1	1	1	1	1
	car2	1	1	1	1	1	1	1	1
	car3	1	1	1	1	1	1	1	1
	car4	1	1	1	1	1	1	1	1
	car5	1	1	1	1	1	1	1	1
	car6	1	1	1	1	1	1	1	1
	person1	0.9801	1	0.8315	0.9079	0.9987	0.9891	1	0.9945
	person2	0.9840	1	0.8644	0.9273	0.9978	1	0.9816	0.9907
	person3	0.9637	1	0.6705	0.8028	0.9844	1	0.8588	0.9240
personx	0.9969	0.125	1	0.2222	0.9969	0.125	1	0.2222	
Traffic Light	1	1	1	1	1	1	1	1	
Traffic Light1	1	1	1	1	1	1	1	1	
		Overall Accuracy = 0.7767				Overall Accuracy = 0.9823			
SUBWAY	person1	1	1	1	1	1	1	1	1
	person2	1	1	1	1	1	1	1	1
	person3	1	1	1	1	1	1	1	1
	person4	0.9821	1	0.8897	0.9416	0.9940	1	0.9632	0.9812
	person5	0.9893	1	0.9338	0.9657	1	1	1	1
	person6	0.9881	1	0.9264	0.9618	1	1	1	1
	person7	0.9916	1	0.9204	0.9585	0.9964	1	0.9659	0.9826
	personx	0.9513	0.023	1	0.0465	0.9905	0.1111	1	0.1999
		Overall Accuracy = 0.9513				Overall Accuracy = 0.9904			
KITESURF	apple	1	1	1	1	1	1	1	1
	person	0.9752	1	0.9404	0.9693	0.9950	1	0.9880	0.9940
	surfboard	0.8811	1	0.7142	0.8333	0.9504	1	0.8809	0.9367
	sportsball	0.9603	1	0.7333	0.8461	0.9851	1	0.9	0.9473
	boat	0.9900	0.3333	1	0.5	0.9900	0.3333	1	0.5
	tennis racket	0.9702	0.1428	1	0.25	0.9752	0.1666	1	0.2857
			Overall Accuracy = 0.8168				Overall Accuracy = 0.9306		
OWN	Bus	0.9703	0.9090	0.5263	0.6666	0.9940	0.9473	0.9473	0.9473
	Busx	0.9861	0.0666	1.0	0.125	0.9980	0.3333	1.0	0.5
	Car1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Car2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Car3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Car4	0.9861	1.0	0.9234	0.9602	0.9980	1.0	0.9890	0.9945
	Carx	0.9111	0.0109	1.0	0.021	0.9861	0.0666	1.0	0.1250
	Truck1	0.9842	0.9220	0.8765	0.8987	0.9842	0.9220	0.8765	0.8987
	Truck2	0.9368	1.0	0.2	0.3333	0.9891	1.0	0.8625	0.9261
	Truckx	0.9980	0.3333	1.0	0.5	0.9990	0.5	1.0	0.6666
		Overall Accuracy = 0.8864				Overall Accuracy = 0.9743			

the observation table it is also clear that for still objects, parameters remain constant.

Table IV gives the details of comparison of our algorithm with others. YOLOv6 is a hardware efficient single stage object detection model comfortable for industrial applications, with hardware-friendly efficient design and high performance. It outperforms YOLOv5 in detection accuracy and inference speed, making it the best OS version of YOLO architecture for production applications. RCNN is double stage detection method based on region based classifier which gives accurate results but it requires many iterations hence it is not suitable for real time applications.

TABLE IV
COMPARISON CHART

OVERALL ACCURACY					
Dataset	RCNN	YOLOv3	YOLOv5	YOLOv6	Modified YOLO
HUMAN	0.7712	0.7767	0.7982	0.8230	0.9823
SUBWAY	0.9541	0.9513	0.9481	0.9564	0.9904
KITESURF	0.7984	0.8168	0.8064	0.8423	0.9306
OWN	0.8435	0.8864	0.8657	0.8989	0.9743

A. Lessons Learned

The proposed work specifically aimed to limit the sudden change in confidence score. We have achieved the success to modify the confidence scores at these places but the exact reasons are still not known. In future work we can focus on various reasons for class change or missing objects. As per class and average speed of the vehicle, in our dataset we can design classified regression module for future speed prediction. In various applications of tracking, the benefits of the deep learning module can be utilized. Linear regression using the machine learning module, can predict the speed of the vehicle from the features extracted using the deep learning module.

B. Limitations

- The proposed solution processes a fix number of subsequent frames to improve the confidence score for classifying an object in the current frame. However, if the confidence score is poor in all the subsequent frames due to variation in illumination or continuous occlusion, the object can still be misclassified by the proposed approach.
- As the proposed approach processes a fixed number of frames (instead of one frame at a time in YOLO) to classify an object in a frame, the time and computation required is relatively high. This may affect the performance of a few applications which requires hard-real-time response. We need to emphasize that the proposed solution can be used without any problem in the near-real time applications such as vehicle counting and density estimation at a junction.
- The proposed solution uses the threshold value of 0.5 for differentiating the inliers and outliers. The value has been tuned to work well with the used datasets under majority of conditions. The proposal can be generalized and extended to customize the threshold automatically for each scenario or dataset in a future work.

V. CONCLUSION

In this work, a modified YOLO tracker which improves the true positive rate is implemented. Here the RANSAC algorithm along with linear interpolation is used by discarding the low confidence detected objects as outlier points and high confidence score points as inlier points. Object detection plays a crucial role in tracking, YOLOv3 is used for object detection which has benefit of accuracy as well as good processing speed. Since the maximum value of the confidence score is 1, the addition of confidence score and classification inaccuracy is 1. To fix classification inaccuracy, centroid coordinate, area of extracted object and confidence score are used. Collection of class probability when compared with mis classified data will automatically confirm the real class of the target object. Classification inaccuracy can be identified only during video processing, as the class change can not be recognized by a single frame. Tracking through the confidence score will give outlier and inlier points. Outlier points solve classification inaccuracy and inlier points give approximate tracking trajectory with the RANSAC algorithm.

Linear interpolation updates outliers as well missing data for a smooth curve. Multiple features from object detection with the YOLOv3 algorithm gives a strong background for processing misclassified data. Implementation of the proposed algorithm on different datasets increases average counting and classification accuracy by 97%.

REFERENCES

- [1] M. B. Khalkhali, A. Vahedian, and H. S. Yazdi, "Vehicle tracking with kalman filter using online situation assessment," *Robotics and Autonomous Systems*, vol. 131, p. 103596, 2020.
- [2] S. Liu and Y. Feng, "Real-time fast moving object tracking in severely degraded videos captured by unmanned aerial vehicle," *International Journal of Advanced Robotic Systems*, vol. 15, no. 1, p. 1729881418759108, 2018.
- [3] N. H. Dardas and N. D. Georganas, "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 11, pp. 3592–3607, 2011.
- [4] R. H. Bhalerao, V. Kshirsagar, and M. Raval, "Finger tracking based tabla syllable transcription," in *Asian Conference on Pattern Recognition*, pp. 569–579, Springer, 2019.
- [5] X. Bian, G. Li, C. Wang, W. Liu, X. Lin, Z. Chen, M. Cheung, and X. Luo, "A deep learning model for detection and tracking in high-throughput images of organoid," *Computers in Biology and Medicine*, p. 104490, 2021.
- [6] 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)*, pp. 1–6, IEEE, 2019.
- [7] Y. Fang, C. Wang, W. Yao, X. Zhao, H. Zhao, and H. Zha, "On-road vehicle tracking using part-based particle filter," *IEEE transactions on intelligent transportation systems*, vol. 20, no. 12, pp. 4538–4552, 2019.
- [8] V. Kshirsagar-Deshpande, T. Patel, A. Abbas, K. Bhatt, R. Bhalerao, and J. Shah, "Vehicle tracking using morphological properties for traffic modelling," in *2020 IEEE India Geoscience and Remote Sensing Symposium (InGARSS)*, pp. 98–101, IEEE, 2020.
- [9] P. C. Niedfeldt and R. W. Beard, "Multiple target tracking using recursive ransac," in *2014 American Control Conference*, pp. 3393–3398, IEEE, 2014.
- [10] J. Chen and L. Dai, "Research on vehicle detection and tracking algorithm for intelligent driving," in *2019 International Conference on Smart Grid and Electrical Automation (ICSGEA)*, pp. 312–315, IEEE, 2019.
- [11] A. R. Pathak, M. Pandey, and S. Rautaray, "Application of deep learning for object detection," *Procedia computer science*, vol. 132, pp. 1706–1717, 2018.
- [12] S. Jha, C. Seo, E. Yang, and G. P. Joshi, "Real time object detection and trackingsystem for video surveillance system," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3981–3996, 2021.
- [13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.
- [14] R. Ghosh, "A modified yolo model for on-road vehicle detection in varying weather conditions," in *Intelligent Computing and Communication Systems*, pp. 45–54, Springer, 2021.
- [15] S. Yang, C. Bo, J. Zhang, and M. Wang, "Vehicle logo detection based on modified yolov2," in *2nd EAI International Conference on Robotic Sensor Networks*, pp. 75–86, Springer, 2020.
- [16] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 502–511, 2019.
- [17] N. Kamel Benamara, E. Zigh, T. Boudghene Stambouli, and M. Keche, "Towards a robust thermal-visible heterogeneous face recognition approach based on a cycle generative adversarial network," 2022.
- [18] B. Ku, K. Kim, and J. Jeong, "Real-time isr-yolov4 based small object detection for safe shop floor in smart factories," *Electronics*, vol. 11, no. 15, p. 2348, 2022.
- [19] S.-H. Chen, C.-W. Wang, I. Tai, K.-P. Weng, Y.-H. Chen, K.-S. Hsieh, *et al.*, "Modified yolov4-densenet algorithm for detection of ventricular septal defects in ultrasound images," 2021.

- [20] W. Han, L. Cao, and S. Xu, "A method of the coverage ratio of street trees based on deep learning," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 7, no. 5, 2022.
- [21] Z. Qiu, Z. Zhao, S. Chen, J. Zeng, Y. Huang, and B. Xiang, "Application of an improved yolov5 algorithm in real-time detection of foreign objects by ground penetrating radar," *Remote Sensing*, vol. 14, no. 8, p. 1895, 2022.
- [22] J. G. A. Barbedo, "A review on methods for automatic counting of objects in digital images," *IEEE Latin America Transactions*, vol. 10, no. 5, pp. 2112–2124, 2012.
- [23] S. Cervantes, A. Mexicano, J.-A. Cervantes, R. Rodríguez, and J. Fuentes-Pacheco, "Binary pattern descriptors for scene classification," *IEEE Latin America Transactions*, vol. 18, no. 01, pp. 83–91, 2020.
- [24] X. Chen, H. Wu, X. Li, X. Luo, and T. Qiu, "Real-time visual object tracking via camshift-based robust framework," *International Journal of Fuzzy Systems*, vol. 14, no. 2, 2012.
- [25] S. Kumar and J. S. Yadav, "Video object extraction and its tracking using background subtraction in complex environments," *Perspectives in Science*, vol. 8, pp. 317–322, 2016.
- [26] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," 2013.
- [27] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2147–2154, 2014.
- [28] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
- [29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.
- [30] Z. Song, Y. Zhang, Y. Liu, K. Yang, and M. Sun, "Msfyolo: Feature fusion-based detection for small objects," *IEEE Latin America Transactions*, vol. 20, no. 5, pp. 823–830, 2022.
- [31] C. J. Veenman, M. J. Reinders, and E. Backer, "Resolving motion correspondence for densely moving points," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 54–72, 2001.
- [32] K. Shafique and M. Shah, "A noniterative greedy algorithm for multi-frame point correspondence," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 1, pp. 51–65, 2005.
- [33] A. S. Silva, F. M. Q. Severgnini, M. L. Oliveira, V. M. S. Mendes, and Z. A. Peixoto, "Object tracking by color and active contour models segmentation," *IEEE Latin America Transactions*, vol. 14, no. 3, pp. 1488–1493, 2016.
- [34] R. Xia, Y. Chen, and B. Ren, "Improved anti-occlusion object tracking algorithm using unscented rauch-tung-strieber smoother and kernel correlation filter," *Journal of King Saud University-Computer and Information Sciences*, 2022.
- [35] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, pp. 13–es, 2006.
- [36] M. Mandal and S. K. Vipparthi, "An empirical review of deep learning frameworks for change detection: Model design, experimental frameworks, challenges and research needs," *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [37] M. Adimoolam, S. Mohan, G. Srivastava, *et al.*, "A novel technique to detect and track multiple objects in dynamic video surveillance systems," 2022.
- [38] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022.
- [39] W. Ben, "Real-time photographing and filming detection on mobile devices," B.S. thesis, University of Twente, 2022.



Varsha Kshirsagar Deshpande pursuing Ph.D from IITRAM Ahmedabad, Working as a Assistant Professor and Head in RMD Sinhgad School of Engineering, Pune. She has completed Master of Engineering from VESIT Chembur, Mumbai in 2010. Her research areas are Image Processing, Wireless Sensor Network, Robotics and its application.



Raghavendra Bhalerao holds the position of Assistant Professor, Electrical and Computer Science Engineering department in Institute of Infrastructure Technology Research and Management (IITRAM). He completed M.Tech in Spatial Information Technology from School of Electronics, Devi Ahliya Vishwavidyalaya Indore, in 2010 (Gold Medallist). He received Ph.D.from the Center of Studies in Resources Engineering (CSRE,IIT Bombay) in 2016. His area of research are Tri-Stereo Image Analysis, Digital Image processing, Applications of

IP to Remote Sensing, Medical Image analysis.



Manish Chaturvedi is working as a Assistant Professor, Electrical and Computer Science Engineering department in IITRAM. He completed his MTech and Ph.D. from Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT). His research interests include the design of Intelligent Transportation Systems, Embedded Systems and IoT, Scalable protocol design for large distributed systems, and the application of ICT for solving problems of societal importance. Recently, he developed interest in distributed data structures,

peer to peer content sharing, and Blockchain framework.