https://www.ijsrtm.com
Vol.5 Issue 1 Mar 2024: 13-22
Published online 11 March 2025
E-ISSN: 2583-7141

International Journal of Scientific Research in Technology & Management

Design and Implementation of Visual Object Tracking with Convolutional-RPN and PP-Yolo

Amit Saxena
Computer Science & Engineering
Rabindranath Tagore University
Bhopal, Madhya Pradesh, India
amit.saxena78@gmail.com

Sitesh Kumar Sinha
Computer Science & Engineering
Rabindranath Tagore University
Bhopal, Madhya Pradesh, India
siteshkumarsinha@gmail.com

Sanjeev Kumar Gupta
Computer Science & Engineering
Rabindranath Tagore University
Bhopal, Madhya Pradesh, India
sanjeevgupta73@yahoo.com

Abstract*— Visual tracking is an important exploration point in the field of computer vision technology. In the first frame it is required to put the target as per the size and location of the object in the reference of x and y co-ordinates and keep tracking the object in the upcoming frames till the last one. The aim of Visual Object Tracking is to automatically acquire the environment of the object in the ensuing video outlines. Visual tracking is now very useful for tracking various moving objects like football in a match, basketball, birds and many more for efficiently tracking the target for better decision making. In Artificial Intelligence; visual tracking is more challengeable because of instability of object in the frames. Conventional techniques are not efficient to deal with this kind of challenges. For tracking the object more efficiently; it has been targeted to achieve it through machine learning techniques. The main purpose of the system is to obtain the pattern of the object by object classification method and then after that tracking the object accordingly. System is depend on two distinct approaches first one is PP-Yolo which is an object classification method; it is associated with the tensorflow that helps to classify the object more precisely. Second one is C-RPN which is object tracking approach on the basis of pattern of the objects. Here the system has been challenges with various benchmarks such as Motion Blur, Low Resolution, Background Clutter, In-plane or Out-plane Rotation, Out of the view, Occlusions, Illumination variations, scale variation and fast motion. Yolo has been especially designed for object classification or object identification, it has great potential to identify the object at real time with high level of precision with high score rate. Yolo is bit associated with the tensorflow because tensorflow is the origin of object classification. All are the precompiled library that are generally utilized in python IDLE with better level of optimization. Object discovery is a computer vision strategy in which a product framework can recognize, find, and follow the object from a frame.

Keywords*— Visual Tracking, Object Detection, C-RPN, PP-Yolo, Object Tracking, OTB50, OTB100, Feature Extraction, Pattern Recognition.

I. INTRODUCTION

Visual object tracking is a computer vision technique used to follow and monitor the movement of objects within a video sequence or a series of frames. The goal of object tracking is to maintain the identity and location of a specific object as it moves through the video frames, despite changes in appearance, scale, orientation, lighting conditions, and potential occlusions. Object tracking has a wide range of applications, including surveillance, robotics, autonomous vehicles, augmented reality, and more [1]. In the field of object detection; it is required to detect the region of interest (ROI) and the background area. Several methods for foreground identification which is the target part of the image were put forth by several researchers over the past ten years. However, the issues have drastic shifts and target drift during tracking, which is not solved by several methods. Accurately estimating the object position is the fundamentally difficult specially in moving object identification and tracking. Identifying moving objects is a crucial stage in the field of computer vision technologies. It is utilized in a variety of applications, including military usage, medical imaging, and video analysis. Typically, a frame comprises information about the foregrounds and backgrounds but it is not the worth technique through accurate information can be obtained. The ROI's feature points in this foreground item represent it, while the remaining features are regarded as background. The two main components of a surveillance system are motion estimation and the detection of moving objects [2].

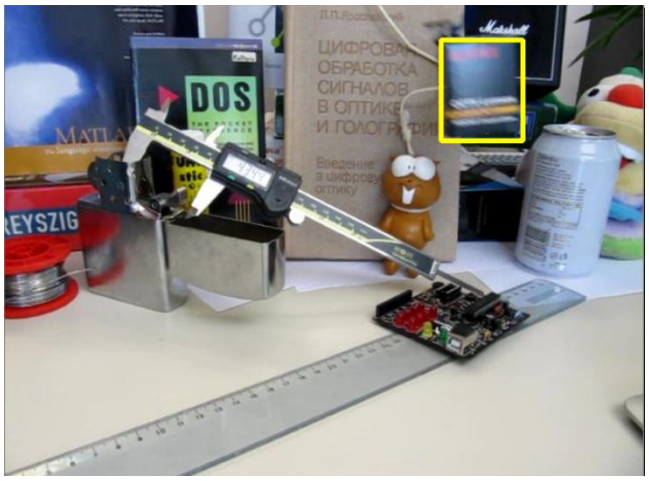


Fig. 1. Box Tracking from TB50 [3]

Fig. 1 shows the system tracking the box which has been obtained from OTB50 benchmark. In the field of object detection; it is required to detect the region of interest (ROI) and the background area. Several methods for foreground identification which is the target part of the image were put forth by several researchers over the past ten years. However, the issues have drastic shifts and target drift during tracking, which is not solved by several methods. Accurately estimating the object position is the fundamentally difficult specially in moving object identification and tracking. Identifying moving objects is a crucial stage in the field of computer vision technologies. It is utilized in a variety of applications, including military usage, medical imaging, and video analysis. Typically, a frame comprises information about the foregrounds and backgrounds but it is not the worth technique through accurate information can be obtained. The ROI's feature points in this foreground item represent it, while the remaining features are regarded as background. The two main components of a surveillance system are motion estimation and the detection of moving objects. The backdrop pixel information affects the object detection, which is the initial phase. Video data must be compressed because it is redundant and out of place in both space and time. The movie may be compressed by minimizing the spatial and temporal information. Many academics have numerous approaches that concentrate on picking up things in a video clip. Many of them combine and intersect many procedures, and many of them employ numerous strategies. The technique known as background subtraction is used to remove the interesting moving item from the frames of a video. The majority of non-stationary backdrop and light fluctuations have an impact on the background subtraction. The optical flow method can really eliminate this flaw, however in congested environments it causes false alarms for tracking systems. The object trackers are impacted by background data in the majority of background subtraction scenarios, although this results in misclassification. Additionally, choosing a reliable classifier is a difficult for improving the accuracy.

II. Literature Review

A. Related Works

Object detection is a significant process for performing computer vision related task. It plays an important role in the field of visual object tracking. If it is talking about the real world then it is a challenging task to detect object with high precision rate because of the mazy in appearance. There are various deep learning approaches through which object can be detected precisely but detecting the moving object and tracking it in every frame is bit more challenging. Moving object detection and tracking have wide variety of application such as border surveillance, road activity detection, forest monitoring and many more. There are various deep learning networks such as CNN, R-CNN, SSD, YOLO and many more, these networks are very much capable to detect the object with large training model. Object tracking is next process after object detection because it is required to target the initial point with target object that has been detected and need to track as per the motion. There are various researches have been done in the field of visual tracking or regression. Most of the systems are based on conventional CNN, SVR, Siamese, 3D CNN and many more but all the systems do not met the desired accuracy as challenges could acquire. Most of the systems are not able to perform well with all the 11 attributes or challenges. Most of the system get failed especially in blur attribute and disappearing challenges. Here in this section a literature survey has been done that stated that the flaws in the various implemented system and which kind of techniques they are using [4]. Haojie Li et al. [5] proposed a framework that is related to the MA-Dual approach that tends to acquire the tracking the object on the basis of their patterns from the datasets. It is a part of 3D CNN approach where system works on the basis of basic structure features of the data or object and try to recognize it according and tracking further for better visual regression.

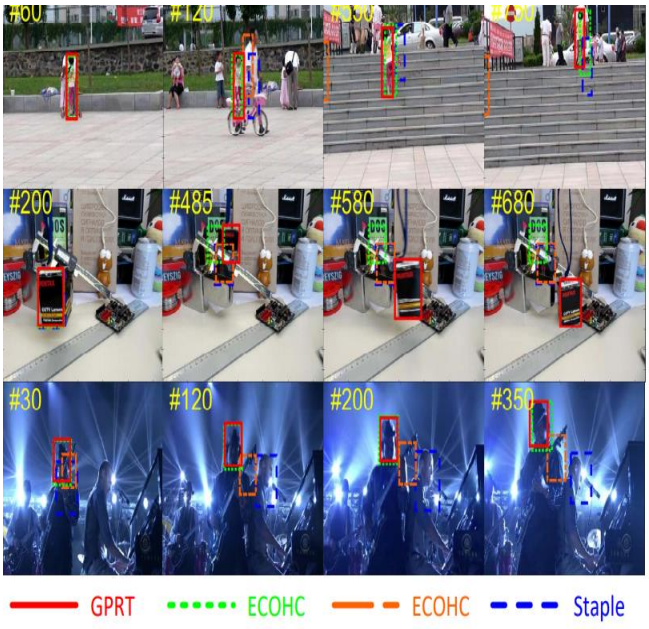


Fig. 2. GPRT Visual Tracking [10]

System is good for certain challenges but not very much efficient for all kind of challenges which have been provided by various benchmarks. Linyu Zheng et al. [6] proposed a GPRT based approach which is bit good as normal tracker. In the comparison with the CF tracking tools the limitations have been removed that has been provided with GPRT. The authors tried to utilize the tracker with updating the network with distinct features with an interval of time. The system has been tested with OTB 2013 as well as OTB 2015 and tried to obtain the accuracy individually; outflanking every one of the current trackers with hand-created highlights. Benchmarks pertain more than 100 videos with thousands of frames in all challenges like birds, bolt, box, cars, bikers, blur bodies, football, human, dudek, david, crowds and many more. Martin Danelljan et al. [7] presented a new approach for dealing with the probability based regression model where system is responsible for tracking objectives on the basis of their physical appearance or thickness of the respective objects. System is intended to track the noisy object too by the help of commotion method. System is based on probabilistic approach where system predicts the target at each frame and tracks it accordingly as per the object movements. It tried to minimize the limitation of the system by reducing the Kullback dissimilarities which is a network based problematic situation that degrade the accuracy of the system by reducing the performance of the tracker. Sangdoo Yun et al. [8] proposed a system which is based on deep reinforcement learning algorithm. In this paper, authors proposed an original activity driven technique utilizing DCNN for visual tracking. System is constrained by an ADNet, that seeks after the objective object by consecutive activities iteratively. The activity driven tracking methodology makes a critical contribution to the decrease of calculation intricacy in tracking. Likewise, RL makes it conceivable to utilize to some degree named information, which could extraordinarily add to the structure of preparing information with a little exertion. As per the assessment results, the proposed tracker accomplishes the best in class execution in 3 casings/s, which is multiple times quicker than the current profound organization based trackers utilizing a tracking-by-location procedure. Besides, the quick form of the proposed tracker accomplishes an ongoing rate (15 Frames/second) by changing the metaparameters of the ADNet, with a precision that outflanks cutting edge constant trackers.

III. Problem Identification

Kai Chen et al. [9] presented an approach that is related to the conventional CNN or traditional convolutional neural network. In this model, network has been trained with various samples or datasets for identifying the pattern of the object where back propagation has also been used to improve the model of the hidden layers. There are various iterations that system performs and back propagation helps to better predict the result or tracking object for obtaining less overflows. A conventional neural network is bit poor in training because it only deals with the trained or well known objects only, it may easily get distracted from the unknown objects. So that is why its pattern recognition is bit poor where system only observe the appearance of the object and once the tracker or bounding box distracted from the region of interest then system become failed to retain the object. System has been trained on the basis of various challenges by using various layers of CNN where hidden layer trying to being with the target object by developing the sensitivity of the system along with increasing the samples of the network that also increases the cost of the system because of large database required for the same. Conventional Convolutional Neural Networks (CNNs) have been widely used and have shown impressive performance in various computer vision tasks. However, they also come with certain disadvantages: Limited Receptive Field: Conventional CNNs with a small number of layers might have a limited receptive field, making it challenging to capture global context and long-range dependencies in an image. Vanishing and Exploding Gradients: As the depth of the network increases, conventional CNNs can suffer from vanishing or exploding gradient problems, leading to difficulties in training deep models effectively. Large Number of Parameters: Deep conventional CNNs can have a large number of parameters, leading to high memory consumption and computational requirements for both training and inference. While conventional CNNs have these limitations, it's important to recognize that they have been foundational in advancing the field of deep learning and computer vision. Many of the disadvantages mentioned here have driven research and development efforts that have led to more advanced architectures with improved performance and efficiency.

IV. Proposed Work & Implementation

The intension of this research proposal is to obtain the object classification method along with object tracking by using C-RPN (Convolutional Region Proposed Network) and PP-Yolo respectively. In this system, there are two kinds of processes have been targeting, first of all system is targeting the object classification for better pattern recognition of an object to track effectively and second one is to purely target the object with bounding box and track accordingly as per the motion or scaling of the object. Both the techniques are based on machine learning approaches with precompiled library for object classification and system is intended to train the network for better visual tracking on the basis of their pattern. System will be examined through OTB50 and OTB100 benchmarks and acquire better level of precision as compare to the base paper. The correlation is processed on both the arrangement branch and the regression branch:

$$A^{cls}_{wxhx2k} = [\phi(x)]_{cls} \star [\phi(z)]_{\text{cls}}$$
$$A^{reg}_{wxhx4k} = [\phi(x)]_{reg} \star [\phi(z)]_{\text{reg}}$$

There are different approaches proposed by various researchers that having certain flaws. The motive of the research is to deep analyze the researches related to the moving object detection along with object tracking. In the discipline of computer vision, object detection is a fundamental study area. For more difficult computer vision

tasks like target tracking, pattern recognition, and semantic comprehension, it serves as a crucial precursor. It seeks to properly identify the category, locate the target of interest inside the image, and provide the bounding box of each target. Small data size, low portability, lack of pertinence, high temporal complexity, and lack of robustness only in certain basic situations are the primary drawbacks of this technology.

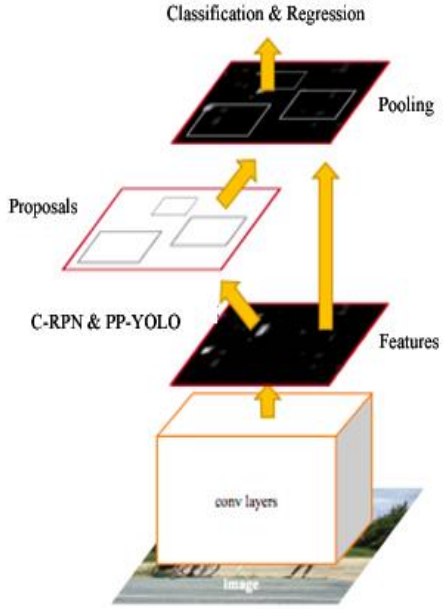


Fig. 3. Regression and Classification using C-RPN and PP-Yolo

Fig. 4 shows the network for detecting an object and classifying the pattern for tracking the target object in further upcoming frames with track score as frame rate as well as with bounding box.

A. C-RPN

A Region Proposal Network (RPN) is a component introduced in the Faster R-CNN (Region Convolutional Neural Network) framework for object detection. It generates region proposals that are potential object bounding boxes in an image. These proposals are then used by the subsequent detection network to classify objects and refine their bounding box positions. If a researcher is looking for advantages that could be related to object tracking, you might be referring to certain techniques or components used in tracking algorithms. One possible concept you might be interested in is "Siamese Networks" or similar architectures. These networks are often used in visual object tracking to learn a similarity metric between target and candidate regions. The Region Proposal Network (RPN) is a crucial component of the Faster R-CNN (Region Convolutional Neural Network) architecture for object detection. It's responsible for generating region proposals that serve as potential object bounding boxes. The RPN operates on the feature maps produced by the backbone network and produces bounding box proposals that are then refined by the subsequent stages of the Faster R-CNN pipeline. The architecture of the RPN is relatively simple and is primarily responsible for generating a diverse set of potential object proposals from the feature maps. These proposals are then used in subsequent stages of Faster R-CNN for object classification and bounding box refinement. Here's a high-level overview of the architecture of the RPN:

a) Input Features: The RPN takes the feature maps obtained from the backbone network as input. These feature maps encode hierarchical information about different image regions at various scales.
b) Convolutional Layers: The RPN consists of a series of convolutional layers that process the input feature maps. These layers are used to extract meaningful features from the feature maps.
c) Anchors: The RPN uses anchor boxes as predefined bounding box templates. These anchor boxes are centered at different positions and have various aspect ratios and scales. The purpose of anchor boxes is to cover a range of possible object sizes and shapes.
d) Convolutional Filter: The convolutional layers in the RPN use a small kernel size (usually 3x3) to slide over the feature maps. At each spatial position, the convolutional filter predicts two types of outputs for each anchor box: objectness scores (indicating whether an object is present or not) and bounding box regressions (adjustments to transform the anchor box into a more accurate bounding box proposal).
e) Objectness Scores: The objectness score represents the probability of the anchor box containing an object. This score helps the RPN determine which anchor boxes are more likely to correspond to objects.
f) Bounding Box Regressions: The bounding box regression outputs provide adjustments to the anchor box dimensions and position to better fit the object within the image.

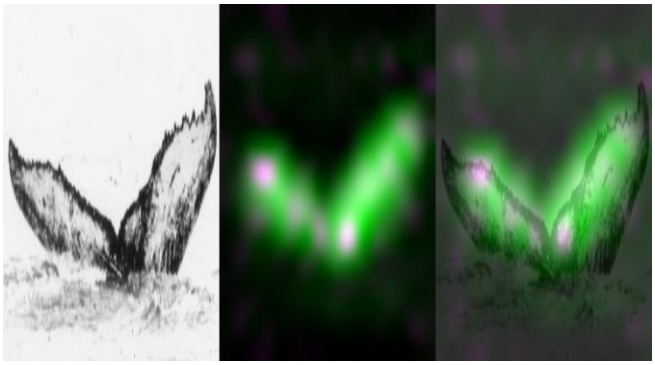


Fig. 4. C-RPN Visualization for Grayscale image

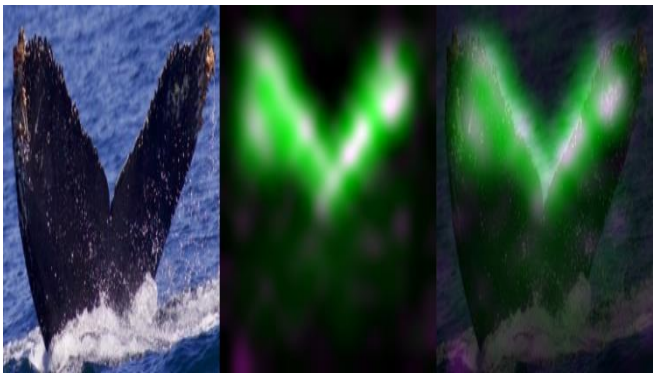


Fig. 5. C-RPN Visualization for RGB image

It has also the potential to track various object simultaneously with multiple boxes. Here the method returns the score for each identification and on the basis of score the prediction level or probabilistic analysis has been introduced. C-RPN is a 4-K localization method where various layers work simultaneously for recognizing the pattern of an object and tracking the object in the entire upcoming frames. It has great potential for obtaining the higher frame rate per second upto 80 frames per sec. In the context of visual object tracking, the term "RPN" might refer to a "Region Proposal Network" or a similar concept. However, as of my last update in September 2021, the primary usage of Region Proposal Networks is in object detection, and their direct application in visual object tracking might not be a common practice.

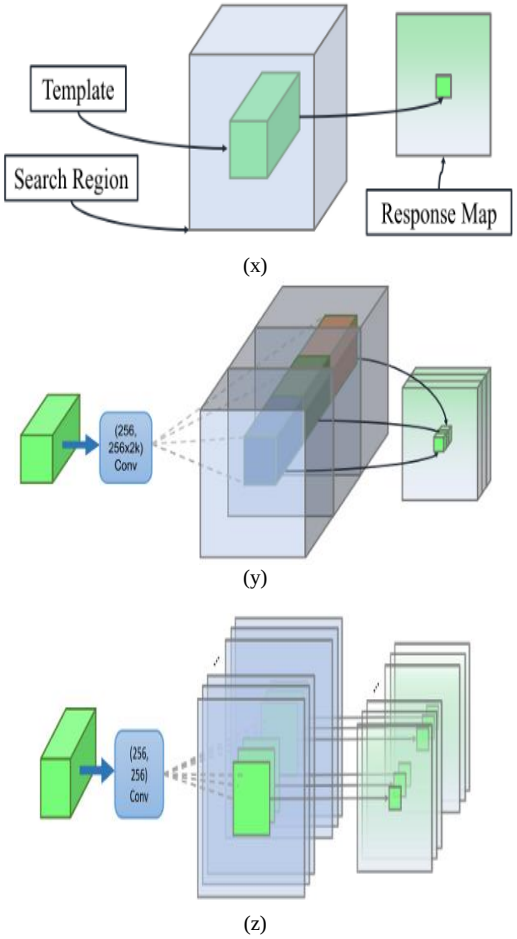


Fig. 6. x,y,z are cross-correlational process for FC, RPN & Yolov3 resp.

B. Yolo

4.3 PP-Yolo

PP-Yolo tends for You Only Look Once which is machine learning model for object classification as well as object identification. It was first depicted in 2015 and now it becomes the best tool for object classification. Yolo is very much associated with the tensorflow network because tensorflow has very large datasets for identifying the objects. All of the image processing based approaches are directly associated with the computer vision library i.e. OpenCV which tends for Open Source Computer Vision. All of the object classification or object detection tools or libraries are uses the precompiled files of OpenCV for optimizing the network for offline uses. Yolo is the new emerging method where system can deal with the real time data and decision can be taken within few seconds with higher accuracy rate and less false recognition rate. System keep checking the object from the frames whether frames pertain the information or not.

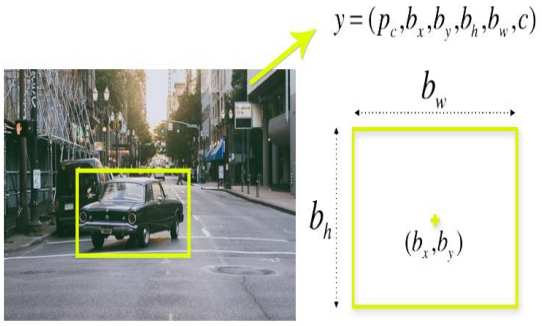


Fig. 7. Yolo Object Detection & Tracking

C. Flow Chart

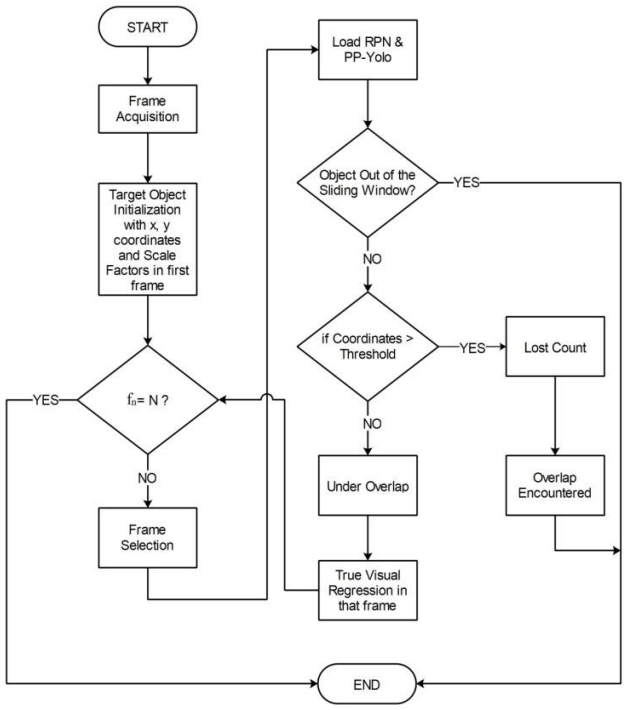


Fig. 8. Flow Chart of Proposed Work

In the very first stage system acquires the frame for preprocessing the model for enhanced input. Once all kind of preprocessing has been done then networks will be fetched and loaded for detaining the files to classify the object on the basis of their appearance. Then another network will be loaded for tracking the object effectively. Here the system uses two certain approaches which are best for recognizing the object and tracking accordingly for better regression model. The efficiency of the system is high and the frame rate per second get enhanced with less overlap counts.

D. RPN and PP-YOLO Algorithm

RPN and PP-YOLO Algorithm

Initialization
Input: Frames
Output: (x, y) Co-ordinates
Step 1: Input the first frame of the dataset
Step 2: Convert RGB to Grayscale image
Step 3: Initialize the target area or object as per the dataset
Step 4: Frame sequences of a dataset $\{X_t\}^T$, X_1 is the first frame
Step 5: Load RPN and PP-Yolo models
Step 6: for *I = 1 to N* **do**
Search target object in frames
if X > Threshold Value **then**
count Overlap++;
else
count True Regression++;
end else
end if
Step 7: Extract coordinates as *Cx, Cy, X, Y*
Where C_x& C_y are the width and height and X, Y are X and Y co-ordinates respectively of the target object.
Step 8: Compare pertained coordinates with the groundtruth
Step 9: Compute accuracy of the system using loss count and true regression
Step 10: End

V. RESULT ANALYSIS

Result is based on co-ordinates which have been obtained from proposed system as compare to the ground truth values. The system which has been proposed will be examined through two benchmarks OTB50 as well as OTB100. Both the benchmarks have numerous videos along with thousands of frames. Each video is related to the certain challenges that may be associated with the benchmark challenges. There are generally eleven attributes for obtaining the accuracy of the system. The system will be evaluated on the basis of overflows as well as the correct recognition of the object at each frame. There will be a threshold value or pixel, if the bounding boxes will slide outside the window then error will encounter and system degraded the accuracy rate. Table No. I shows the performance of benchmark OTB50 and as per all the challenges; the mean accuracy is recorded as 91.72 %. The accuracy which has been acquired by the proposed system is bit higher than the earlier proposed system till now. System pertains minimal error rate with less overlap count.

Table No. I Recorded Accuracy of Benchmark OTB50

Datasets	*Accuracy*	*Datasets*	*Accuracy*
Basketball	100	Human3	100
Biker	100	Human4	54.51761
Bird1	46.42857	Human6	93.14721
BlurBody	99.3994	Human9	100
BlurCar2	85.61644	Ironman	80.76923
BlurFace	64.41718	Jump	96.69421
BlurOwl	98.57434	Jumping	100
Bolt	100	Liqour	59.9537
Box	99.04514	Matrix	89.79592
Car1	100	MotorRolling	95
Car4	100	Panda	100
CarDark	100	RedTeam	100
CarScale	96.8	Shaking	100
ClifBar	96.39066	Singer2	81.0585
Couple	100	Skating1	89.13043
Crowds	100	Skating2	72.12766
David	81.08696	Skiing	100
Deer	100	Soccer	92.59259
Diving	98.59813	Surfur	100
DragonBaby	89.18919	Sylvester	100
Dudek	99.56294	Tiger2	87.32782
Football	55.1532	Trellis	100
Freeman4	94.30605	Walking	100
Girl	100	Walking2	100
		Woman	97.64706
Mean		**91.72102327**	

Graph No. I Recorded Accuracy of Benchmark OTB50

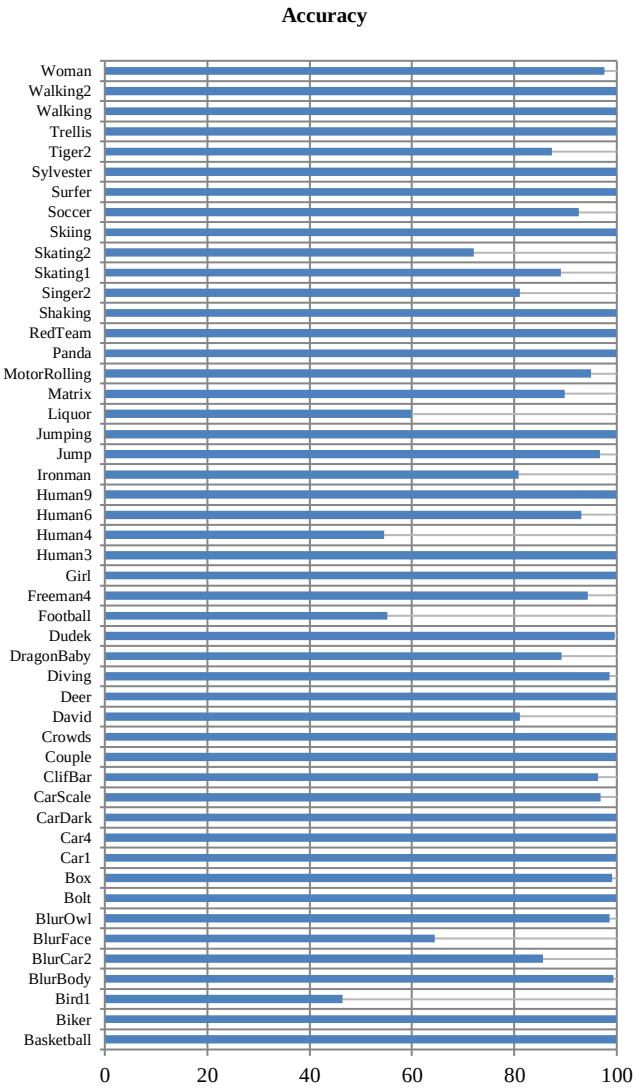


Table No. II Recorded Overlap/Lost of Benchmark OTB50

Datasets	*Overlap*	*Datasets*	*Overlap*
Basketball	0	Human3	0
Biker	0	Human4	45.48239

Bird1	53.57143	Human6	6.85279
BlurBody	0.6006	Human9	0
BlurCar2	14.38356	Ironman	19.23077
BlurFace	35.58282	Jump	3.30579
BlurOwl	1.42566	Jumping	0
Bolt	0	Liqour	40.0463
Box	0.95486	Matrix	10.20408
Car1	0	MotorRolling	5
Car4	0	Panda	0
CarDark	0	RedTeam	0
CarScale	3.2	Shaking	0
ClifBar	3.60934	Singer2	18.9415
Couple	0	Skating1	10.86957
Crowds	0	Skating2	27.87234
David	18.91304	Skiing	0
Deer	0	Soccer	7.40741
Diving	1.40187	Surfur	0
DragonBaby	10.81081	Sylvester	0
Dudek	0.43706	Tiger2	12.67218
Football	44.8468	Trellis	0
Freeman4	5.69395	Walking	0
Girl	0	Walking2	0
		Woman	2.35294
Mean		**8.278976735**	

Graph No. II Recorded Overlap/Lost of Benchmark OTB50

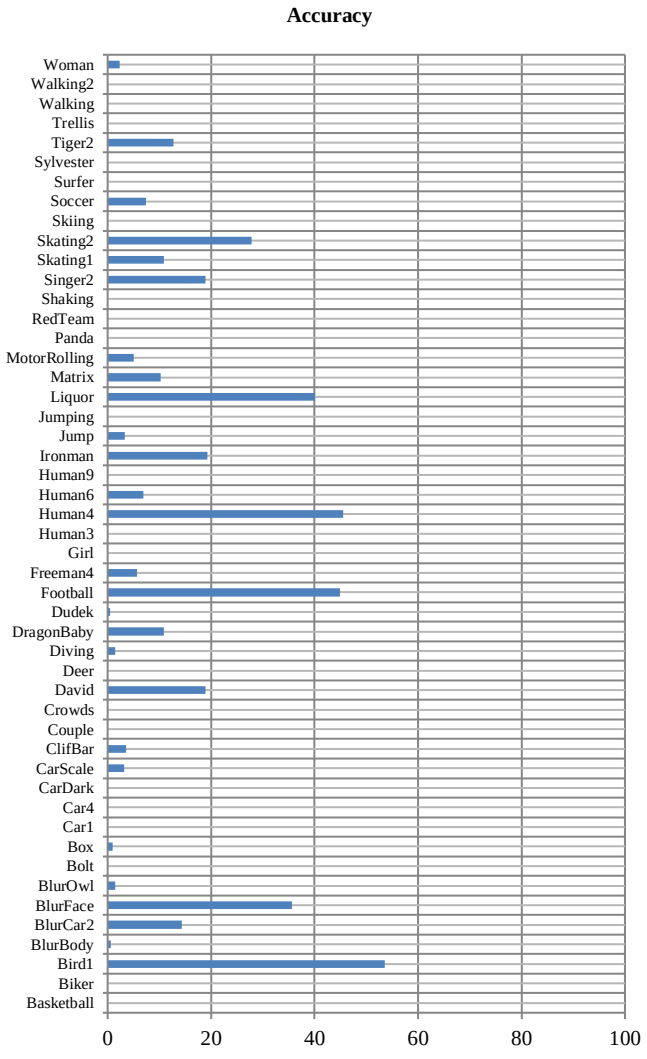


Table No. III Recorded Accuracy of Benchmark OTB100

Datasets	*Accuracy*	*Datasets*	*Accuracy*
Bird2	100	Freeman1	100
BlurCar1	99.05149	Freeman3	100
BlurCar3	100	Girl2	46.30872
BlurCar4	93.61702	Gym	100
Board	60.17442	Human2	94.66667
Bolt2	69.09722	Human5	100
Boy	100	Human7	100
Car2	100	Human8	100
Car24	100	Jogging	100
Coke	93.75	KiteSurf	100
Coupon	41.25	Lemming	93.54354
Crossing	100	Man	100
Dancer	100	Mhyang	100
Dancer2	100	MountainBike	100
David2	100	Rubik	97.59157
David3	98.79032	Singer1	100
Dog	87.5	Skater	100
Dog1	41.01979	Skater2	96.05568
Doll	97.59566	Subway	100
FaceOcc1	41.82638	Suv	100
FaceOcc2	76.72316	Tiger1	99.42529
Fish	100	Toy	98.51301
Fleetface	66.09687	Trans	50
Football1	100	Twinnings	99.78541
		Vase	88.76404
Mean		**90.43155633**	

Graph No. III Recorded Accuracy of Benchmark OTB100

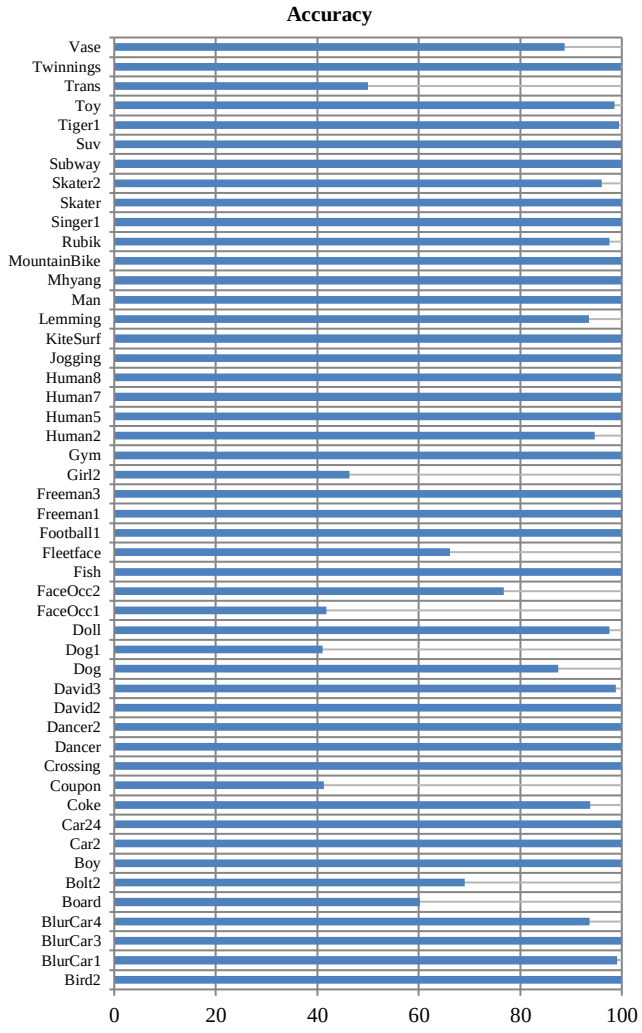


Table No. IV Recorded Overlap/Lost of Benchmark OTB100

Datasets	*Overlap*	*Datasets*	*Overlap*
Bird2	0	Freeman1	0
BlurCar1	0.94851	Freeman3	0
BlurCar3	0	Girl2	53.69128
BlurCar4	6.38298	Gym	0
Board	39.82558	Human2	5.33333
Bolt2	30.90278	Human5	0
Boy	0	Human7	0
Car2	0	Human8	0
Car24	0	Jogging	0
Coke	6.25	KiteSurf	0
Coupon	58.75	Lemming	6.45646
Crossing	0	Man	0
Dancer	0	Mhyang	0
Dancer2	0	MountainBike	0
David2	0	Rubik	2.40843
David3	1.20968	Singer1	0
Dog	12.5	Skater	0
Dog1	58.98021	Skater2	3.94432
Doll	2.40434	Subway	0
FaceOcc1	58.17362	Suv	0
FaceOcc2	23.27684	Tiger1	0.57471
Fish	0	Toy	1.48699
Fleetface	33.90313	Trans	50
Football1	0	Twinnings	0.21459
	0	Vase	11.23596
Mean			**9.568443673**

Graph No. IV Recorded Overlap/Lost Count of Benchmark OTB100

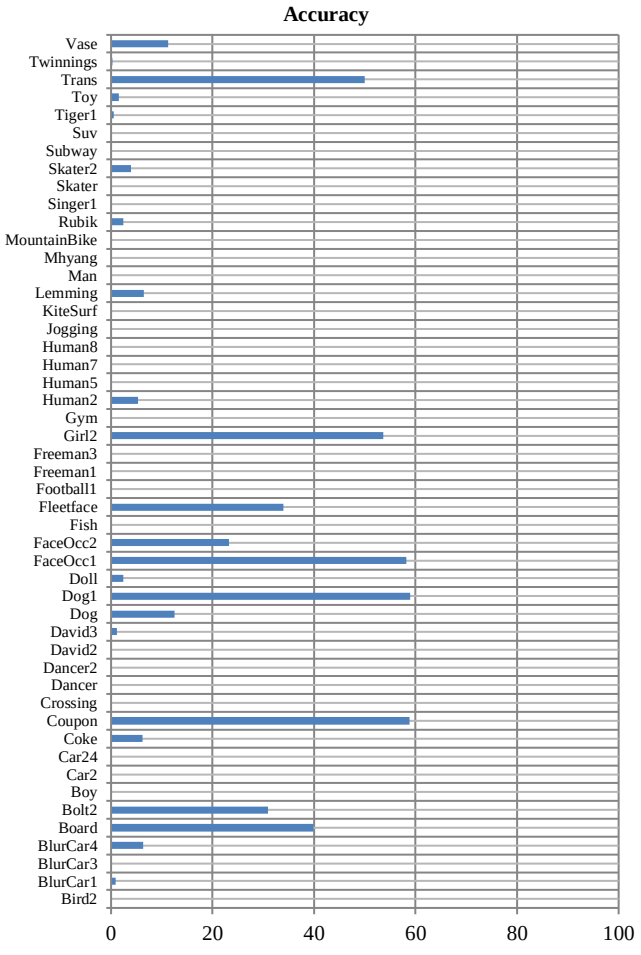


Table No. 1 shows the Overlap/Lost of benchmark OTB100 and as per all the challenges; the mean overlap is recorded as 9.57 %. The mean accuracy of TB50 and TB100 are 91.72 % and 90.43 % respectively. The mean overlap of TB50 and TB100 are 8.28 % and 9.57 % respectively. The datasets have 100 videos with more than 74 thousand of frames that have been adopted from OTB officials. System has been initiated with target values as (x, y, box-width, box-height) which has been pertained from groundtruth values.

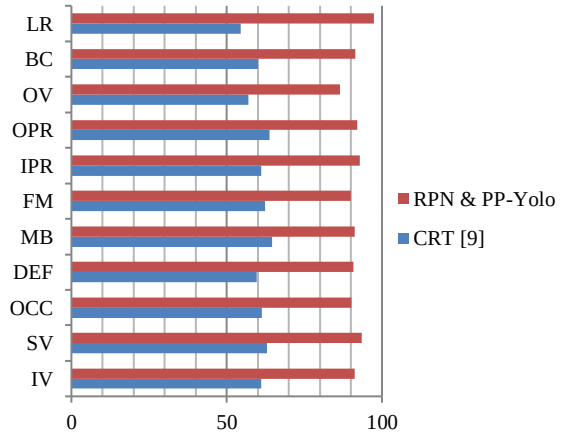
Table No. V Comparison of Evaluations Under 11 Attributes

TB50 - Mean Accuracy in %		
	CRT [9]	**C-RPN & PP-Yolo**
IV	61.1	91.26931636
SV	63.0	93.51202351
OCC	61.3	90.18705643
DEF	59.7	90.76029545
MB	64.7	91.185196
FM	62.4	90.05515846
IPR	61.1	92.95729172
OPR	63.8	92.10928419
OV	57.0	86.52858727
BC	60.1	91.49749
LR	54.5	97.481437

Table No. V Comparison of Evaluations Under 11 Attributes for OTB100

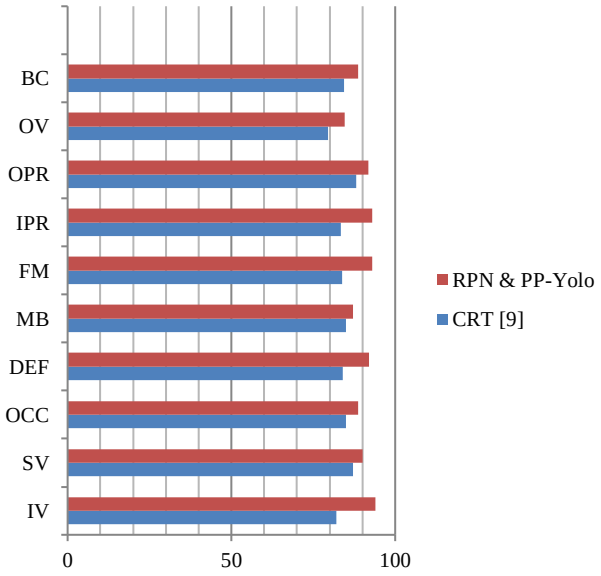
TB100 - Mean Accuracy in %		
	CRT [9]	**C-RPN & PP-Yolo**
IV	82.0	93.90213357
SV	87.1	89.96036857
OCC	84.9	88.75863778
DEF	83.9	91.99110647
MB	85.0	87.21277091
FM	83.7	93.06196
IPR	83.4	92.897175
OPR	88.0	91.79300367
OV	79.5	84.57265333
BC	84.4	88.58849667

Graph No. V Comparison of Evaluations Under 11 Attributes for OTB50



Graph V represents the comparison of accuracies that have been achieved against benchmark TB50 by CRT technique (Previous Work) and Proposed Work respectively. Proposed system pertained bit higher level of accuracy as compare to the earlier proposed system.

Graph No. VI Comparison of Evaluations Under 11 Attributes for OTB100



Graph VI represents the comparison of accuracies that have been achieved against benchmark TB100 by CRT technique (Previous Work) and Proposed Work respectively.

VI. Conclusion & Future Scope

The main objective of the system is to obtain the pattern of the object by object classification method and then tracking the object accordingly. System depends on two distinct approaches first one is Yolov3 which is an object classification method; it is associated with the tensorflow that helps to classify the object more precisely. Second one is C-RPN which is object tracking approach on the basis of pattern of the objects. Here the system has been challenges with various benchmarks such as Motion Blur, Low Resolution, Background Clutter, In-plane or Out-plane Rotation, Out of the view, Occlusions, Illumination variations, scale variation and fast motion. System is targeting to obtain better level of accuracy as compare to the previous one and will be tested for OTB50 and OTB100 benchmarks. System will also pertain less overflow rate for each frame by enhancing the network model of the system that will identify and track the object more precisely. In future system can be tested with VOT2016, VOT2018, TempleColor128 and many more for accepting the challenges and might pertains better level of accuracy as compare to the earlier proposed system. System may also use next version of Yolo like Yolo-v4 or v5 which could be introduced in upcoming years or any other object classification or detection technique for better precision in future. There are so many applications of visual tracking model such as foul detection in football matches and many more. So, the scope of this work is bit demanding in future and accuracy matters a lot. So, in future accuracy may be enhanced and tracking can be more specific and precise.

References

[1] Peter Mountney, Danail Stoyanov & Guang-Zhong Yang (2010). "Three-Dimensional Tissue Deformation Recovery and Tracking: Introducing techniques based on laparoscopic or endoscopic images." IEEE Signal Processing Magazine. 2010 July. Volume: 27" (PDF). IEEE Signal Processing Magazine. 27 (4): 14–24. doi:10.1109/MSP.2010.936728. hdl:10044/1/53740.

[2] Lyudmila Mihaylova, Paul Brasnett, Nishan Canagarajan and David Bull (2007). Object Tracking by Particle Filtering Techniques in Video Sequences; In: Advances and Challenges in Multisensor Data and Information. NATO Security Through Science Series, 8. Netherlands: IOS Press. pp. 260–268. CiteSeerX 10.1.1.60.8510. ISBN 978-1-58603-727-7.

[3] VOT Challenges, Datasets, 2015. [Online]. Available: https://www.votchallenge.net/vot2016/dataset.html, [Accessed: 13-Aug- 2023]

[4] Yang, L.; Zhou, H.; Yuan, G.; Xia, M.; Chen, D.; Shi, Z.; Chen, E. SiamUT: Siamese Unsymmetrical Transformer-like Tracking. Electronics 2023, 12, 3133. https://doi.org/10.3390/electronics12143133

[5] H. Li, S. Wu, S. Huang, K. Lam and X. Xing, "Deep Motion-Appearance Convolutions for Robust Visual Tracking," in IEEE Access, vol. 7, pp. 180451-180466, 2019, doi: 10.1109/ACCESS.2019.2958405.

[6] Linyu Zheng, Ming Tang, Jinqiao Wang, " Learning Robust Gaussian Process Regression for Visual Tracking," in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Main track. Pages 1219-1225. https://doi.org/10.24963/ijcai.2018/170.

[7] Martin Danelljan, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7183-7192

[8] S. Yun, J. Choi, Y. Yoo, K. Yun and J. Y. Choi, "Action-Driven Visual Object Tracking With Deep Reinforcement Learning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 6, pp. 2239-2252, June 2018, doi: 10.1109/TNNLS.2018.2801826.

[9] K. Chen and W. Tao, "Convolutional Regression for Visual Tracking," in IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3611-3620, July 2018, doi: 10.1109/TIP.2018.2819362.

[10] Zhang, Da & Maei, Hamid & Wang, Xin & Wang, Yuan-Fang. (2017). Deep Reinforcement Learning for Visual Object Tracking in Videos.

[11] J. F. Henriques, R. Caseiro, P. Martins and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 583-596, 1 March 2015, doi: 10.1109/TPAMI.2014.2345390.

[12] D. S. Bolme, J. R. Beveridge, B. A. Draper and Y. M. Lui, "Visual object tracking using adaptive correlation filters," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2544-2550, doi: 10.1109/CVPR.2010.5539960.

[13] Dogra, Debi & Badri, Vishal & Majumdar, Arun & Sural, Shamik & Mukherjee, Jayanta & Mukherjee, Suchandra & Singh, Arun. (2014). Video analysis of Hammersmith lateral tilting examination using Kalman filter guided multi-path tracking. Medical & biological engineering & computing. 52. 10.1007/s11517-014-1178-2.

[14] S. Yun, J. Choi, Y. Yoo, K. Yun and J. Y. Choi, "Action-Decision Networks for Visual Tracking with Deep Reinforcement Learning," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1349-1358, doi: 10.1109/CVPR.2017.148.

[15] Linyu Zheng, Ming Tang, and Jinqiao Wang. 2018. Learning robust Gaussian process regression for visual tracking. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18). AAAI Press, 1219–1225.

[16] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing and J. Yan, "SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4277-4286, doi: 10.1109/CVPR.2019.00441.

[17] Xin Li, Qiao Liu, Nana Fan, Zikun Zhou, Zhenyu He, Xiao-yuan Jing, Dual-regression model for visual tracking, Neural Networks, Volume 132, 2020, Pages 364-374, ISSN 0893-6080, https://doi.org/10.1016/j.neunet.2020.09.011.

[18] B. Zhang, X. Zhang and J. Qi, "Support vector regression learning based uncalibrated visual servoing control for 3D motion tracking," 2015 34th Chinese Control Conference (CCC), 2015, pp. 8208-8213, doi: 10.1109/ChiCC.2015.7260942.

[19] T. Wang and W. Zhang, "The visual-based robust model predictive control for two-DOF video tracking system," 2016 Chinese Control and Decision Conference (CCDC), 2016, pp. 3743-3747, doi: 10.1109/CCDC.2016.7531635.

[20] Djelal, N.; Saadia, N.; Ramdane-Cherif, A. (2012). [IEEE 2012 2nd International Conference on Communications, Computing and Control Applications (CCCA) - Marseilles, France (2012.12.6-2012.12.8)] CCCA12 - Target tracking based on SURF and image based visual servoing. , (), 1–5. doi:10.1109/ccca.2012.6417913

[21] C. H. Li and T. I. James Tsay, "Robust Visual Tracking in Cluttered Environment Using an Active Contour Method," 2018 57th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), 2018, pp. 53-58, doi: 10.23919/SICE.2018.8492705.

[22] Q. Guo, W. Feng, R. Gao, Y. Liu and S. Wang, "Exploring the Effects of Blur and Deblurring to Visual Object Tracking," in IEEE Transactions on Image Processing, vol. 30, pp. 1812-1824, 2021, doi: 10.1109/TIP.2020.3045630.

[23] H. Li and Y. W eds., (2015). Object of interest tracking based on visual saliency and feature points matching. International Conference on Wireless, Mobile and Multi-Media, pp. 201-205.

[24] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In ECCV Workshops, 2016.

[25] G. Bhat, J. Johnander, M. Danelljan, F. Shahbaz Khan, and M. Felsberg. Unveiling the power of deep tracking. In ECCV, September 2018. 7

[26] D. Bolme, J. Beveridge, B. Draper, and Y. Lui. Visual object tracking using adaptive correlation filters. In CVPR, 2010. 2

[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018. 2

[28] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In CVPR, 2017.