Volume 66, Issue 1, 2022

Journal of Scientific Research

of

The Banaras Hindu University

Convolutional Siamese-RPN++ and Yolo-v3 based Visual Tracking Regression

Ishita Jain[*1], Sanjay Kumar Sharma[2]

[*1]Electronics & Communication Engineering UIT, RGPV, Bhopal, Madhya Pradesh, India, ishitajain1997@gmail.com
[2]Electronics & Communication Engineering UIT, RGPV, Bhopal, Madhya Pradesh, India, sksharma@rgtu.net

Abstract: **Visual tracking is an implementation of moving object tracing from deep machine learning methods where system initially set the object and generate a unique identification or pattern for tracking the moving object at each frame of a video. Object tracking is the undertaking of automatically distinguishing objects in a video and deciphering them as a bunch of directions with high accuracy. This paper intended to propose a SiamRPN network which has been considered as offline network with having very large dataset. In this network there are so many sub networks are available to extract the features along with regression and classification. Here the Siamese-RPN++ has been reconciled with Yolo-v3 which is an object detection approach that enhances the feature extraction model for better visual tracking analysis. Prior recognition frameworks repurpose the classifiers or localizers to perform feature extraction. It applies the model to an image at various areas even while object scaling. System has been tested with various datasets/benchmarks including OTB50 and OTB100 and achieved 91.17 & 89.98 resp. percent of accuracies.**

Index Terms: **Visual Tracking, Object Detection, Siamese-RPN++, Yolo-v3, Object Tracking, OTB50, OTB100, Feature Extraction, Pattern Recognition.**

I. Introduction

Visual regression is the process of tracking the object location while movements. In terms of various aspects visual regression is bit harder for those objects which are fast in motion. It is bit challenging for human to tract the motion or actual location of the objects while having in high frame rate, so indeed it is often more challenging for human to interact or extract the actual location of the objects with high precision [1]. In the field of visual tracking systems, there are several researches have been attempted and reach their significant roles. They have been tested their system with so many datasets such as OTB50 and OTB100. But do not meet the desired precision and recall. In an object is in visibility in entire frame then system is efficient to track the location of the target object. The network could be particularly troublesome while having in quick motion comparative with frame per second [2]. Another circumstance that expands the intricacy of the issue is the point at which the followed object changes direction after some time. For these circumstances visual tracking frameworks for the most part utilize a movement that portrayed the efficiency of the object. There are so pre-trained networks available that challenges the researchers to track the object with high preciseness. They trained the network on the basis of object's textures and features. System targeted the object on the basis of same and track it in all consequent frames and if object lost the visibility or system distracted from the target object then it is hard to secure the precision [3].

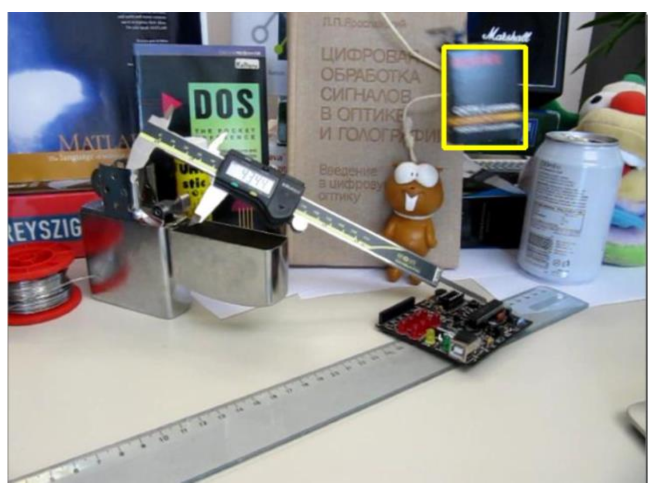


Fig. 1. Box Tracking from TB50

Fig. 1 shows the system tracking the box which has been obtained from OTB50 benchmark. Here the system has been divided into two different categories. In the very first category,

[*]Ishita Jain

system pertains the initial channel to obtain the feature of an object in the very first frame by using Fourier domain. And in the second category, system is intended to tract that extracted feature in all frames by following the Fourier data and extract the local area in the frames [4]. Ongoing correlation channel based strategies utilize profound elements to work on the precision, however it generally hurts the speed during model update [5]. Another part of strategies intends to utilize extremely impressive profound components and don't refresh the model [6]. Be that as it may, in light of the fact that the domain explicit data isn't utilized, the application of this strategy is not efficient and not considered as co relational channel. Paper intended that disconnected prepared profound learning based tracker can accomplish serious outcomes contrasted with the best in class correlation channel based techniques when appropriately planned. The correlations between the object and the patterns should be comparatively same in each and every frame, but it is not possible to remain intact in mind because object may get changed after a particular frame in the respect of shape, size and patterns. Object may get disappear in a particular frame and system may lost the tracking area that also may distract the bounding box that directly degraded the precision of the system and it is bit challenging with different datasets.

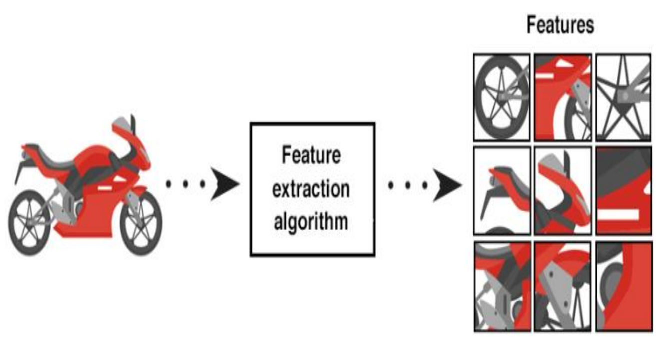


Fig. 2. Visual Represetation of Feature Extraction of an Object [8]

In profound learning, system is able to highlight the object in various frames. System includes feature extraction separation and find an importance aspects to follow the objects at its possible extent. System has various weights and layer to classify the object and highlight to track with local area extraction. Neural based extractors may classifies the object from initial frame to the end frame as opposed to standard ML models that use hand-made provisions [8].

II. Related Works

HAOJIE LI et al. [9] proposed a network that is intended to track the object with MA-Dual technique that is following spatial transient approach for tracking the patterns in each and every frames of a dataset. This paper is based on 3d convolutional approach where system extracts the features on the basis of their structures and follow the same in entire frame. System is bit inefficient to obtain the object location in certain datasets because some challenges are bit difficult because of various prospects such as motion blur, low resolutions and many more. System integrity may get differ accordingly because data may get highlighted distinctly in different luminance. System has been tested with various datasets such as UAV123, OTB benchmarks, VOT and TC128 too. The investigation results show that the proposed strategy accomplishes an exceptionally encouraging tracking execution, and is particularly acceptable at taking care of testing conditions, like disfigurement, scale variety, enlightenment changes, and so forth. Linyu Zheng et al. [10] proposed a Gaussian Process Regression based tracker (GPRT) which is a reasonably normal tracking approach. Contrasted with all the current CF trackers, the limit impact is wiped out completely and the part stunt can be utilized in our GPRT. Also, Authors present two productive and successful update techniques for our GPRT. Analyses are performed on two public datasets: OTB-2013 and OTB-2015. Without extravagant accessories, on these two datasets, our GPRT acquires 84.1% and 79.2% in mean cross-over exactness, individually, outflanking every one of the current trackers with hand-created highlights. An original tracking framework, GPRT which applying the Gaussian Regression Processes to visual tracking, has been introduced in this paper. Contrasted with all the current CF trackers, our GPRT not exclusively doesn't exist the limit impact, however al so can exploit the bit stunt simultaneously. Expansion, Authors propose two distinct proficient and compelling up date techniques for our GPRT. Authors perform extensive tests on two benchmark datasets: OTB-2013 and OTB 2015.

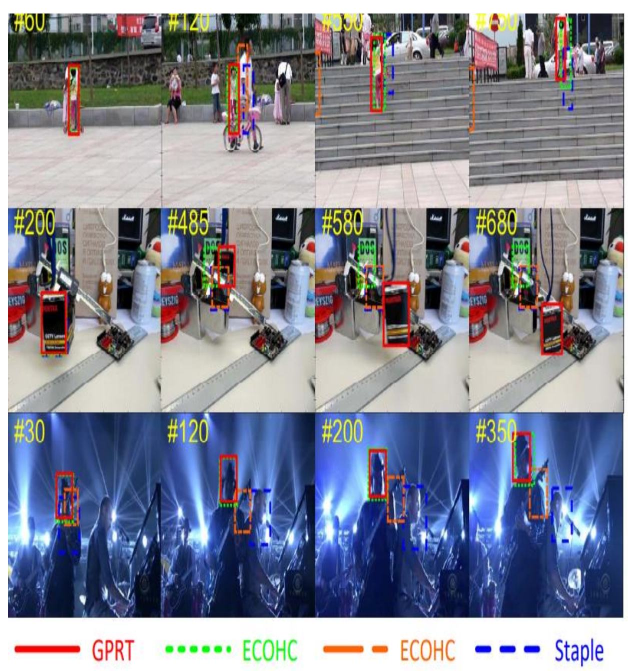


Fig. 3. GPRT Visual Tracking [10]

Martin Danelljan et al. [11] proposed a probabilistic regression detailing and apply it to tracking. System's network predicts the restrictive likelihood thickness of the objective state given an info image. Significantly, system's plan is equipped for demonstrating name commotion originating from in precise explanations and ambiguities in the assignment. The regression network is prepared by limiting the Kullback Leibler dissimilarity. When applied for tracking, system's definition not just permits a probabilistic portrayal of the yield, yet additionally generously works on the presentation. System's tracker sets another best in class on six datasets, accomplishing 59.8% AUC on LaSOT and 75.8% Accomplishment on TrackingNet. Kai Chen et al. [12] proposed a regression method that follows the convolutional network for tracking the moving object. Here the system is based on edge regression that also extract the edges for tracking the object by its patterns and textures. System is also based on back propagation model that follows the rendering technique with different layers of convolutional model. In the DCF model, each layer has been designed or trained with different prospective that follows the various integration features that object pertains and on the basis of these parameters system tracks the object with different proportions and challenges with various iterations and back propagations. It is one more approach to manage come out as comfortable with the relapse model for visual following single convolutional layer. Maybe than learning the immediate relapse model in a shut construction, creators endeavor to deal with the relapse issue by propelling a one-channel-yield convolution layer with GD. In particular, the piece size of the convolution layer is set to the size of the item. Rather than DCF, it is attainable to intertwine all "certified" models cut from the whole picture. An essential issue of the GD approach is that most of the convolutional tests are negative and the responsibility of positive models will be covered. To determine this issue, creators propose a cunning target ability to clear out straightforward negatives and update up-sides. To accelerate the preparation stage, authors additionally propose a worked on objective capacity to kill simple negatives and improve positives. The outcomes show that the proposed calculation accomplishes extraordinary execution and beats the majority of the current DCF-based calculations.

III. Problem Identification

Kai Chen et al. [12] introduced a system which is based on convolutional regression that is conventional CNN. System tracking the object by using a trained network using CNN, but this approach is bit conventional for visual tracking because it uses back-propagation method and back-propagation is a strategy to discover the contribution of each weight in the errors after a group of information is inclined and the majority of good improvement algorithms (SGD, ADAM) utilizes back-propagation to discover the angles, back-propagation has been doing as such great task but somewhat it is certainly not a productive method of learning, since it needs huge dataset. At the point when authors say translational invariance authors imply that a similar object with marginally change of direction or position probably won't start up the neuron that should perceive that object. Pooling layers is a serious mix-up on the grounds that it loses a ton of significant data and it disregards the connection between the part and the entirety. CNN's are magnificent however it has 2 exceptionally risky defects Interpretation invariance and pooling layers, fortunately author can diminish the risk with information increase yet something is coming up (capsule networks). Object detection matters in the field of object tracking because a pattern or feature can effectively analyzed for tracking as compare to the any conventional method. The proposed system uses two different methods and combining them for acquiring better precision.

IV. Proposed Work & Implementation

The aim of the system is to track target object throughout the entire frame with better precision without encountering high overflow. Here system uses Siamese-RPN++ and Yolo-v3 methodologies for tracking object in a video frames with various challenge factors such as Deformations (DEF), Illumination Variations (IV), Background Clutter (BC), In-Plane Rotations (IPR), Fast Motions (FM), Occlusions (OCC), Out of Plane Rotations (OPR), Motion Blurs (MB), Scale Variations (SV), Out of Views (OV) and Low Resolutions (LR). A Siam network comprises of distinct twigs that verifiably encodes the first fixes to another space and afterward combines them with a particular tensor to deliver a solitary yield. It's generally utilized for looking at two branches' provisions in the certainly implanted space particularly for contrastive errands. As of late, Siamese networks have attracted extraordinary consideration visual tracking local area in view of their fair precision and speed. The proposed work is based on Siam network which has been trained for pattern recognition of different object on the basis of their look or patterns. System is regressed with RPN networks and it is a regional proportional network that is able to track the object's location or area on the basis of pattern classification. This network is broad and convolutionally well trained for high featured object patterns as well as low resolution data. Here the system not only uses the Siam network, but rather than that system also uses Yolo v3 based network for object detection and classification that helps to obtain the object correctly. Let $L\tau$ denote the translation operator $(L\tau\ x)[u] = x[u - \tau\ ]$, then all paddings are removed to satisfy the definition of fully convolution with stride k:

$$h(L_{k\tau}x) = L_{\tau}h(x)$$

The two branches share boundaries in network so the two patches are verifiably encoded by a similar change which is reasonable for the resulting errands. For accommodation, we mean $\phi(z)$ and $\phi(x)$ as the yield highlight guides of Siamese subnetwork. The locale proposition subnetwork comprises of a couple shrewd correlation area and a management segment. The supervision segment has two branches, one for forefront foundation arrangement and the other for proposition regression. In case there are k anchors, network needs to yield 2k channels for arrangement and 4k channels for regression. So the pair-wise correlation area first increment the channels of $\phi(z)$ to two branches $[\phi(z)]cls$ and $[\phi(z)]reg$ which have 2k and 4k occasions in channel separately by two convolution layers. $\phi(x)$ is additionally parted into two branches $[\phi(x)]cls$ and $[\phi(x)]reg$ by two convolution layers however keeping the channels unaltered. $[\phi(z)]$ is filled in as the correlation kernel of $[\phi(x)]$ in a "bunch" way, in other words, the divert number in a gathering of $[\phi(z)]$ is equivalent to the by and large channel number of $[\phi(x)]$. The correlation is processed on both the arrangement branch and the regression branch:

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

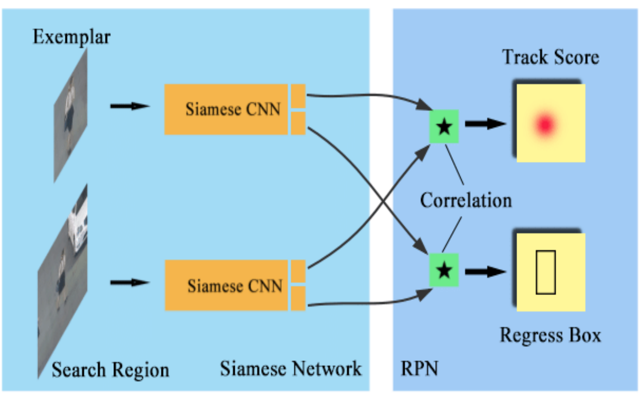


Fig. 4. Siamese Pattern Recognition

Fig. 4 shows the SiameseRPN 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. SiameseRPN++

Visual tracking is a fashion for obtaining the object in a classic manner from video data in the form of sequences of frames and track a target object in entire frames without having information regarding the target object but rather than that it only knows the pattern in the very first frame and on the basis of that it follows that patterns from first frame to the end frame effectively. A video distributes into various number of frames and Siam network has been initialized with target area manually that is also called area of interest or region of interest. It is based on cross correlation model that embedding the technology of local feature extraction with corresponding classes and looking for the proportional region by comparing the corresponding frames with RPN. It is also based on Alexnet feature extractor that produces the templates of features that later compares with the object feature for tracking the model more efficiently.

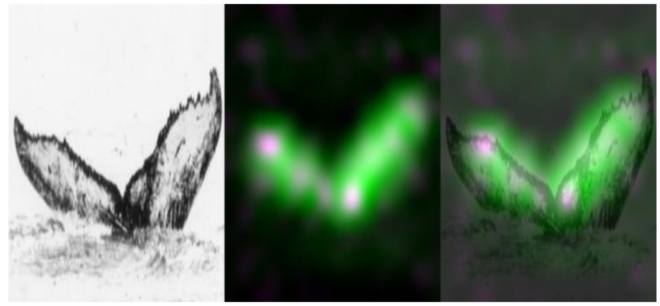


Fig. 5. SiameseRPN++ Visualization from Grayscale Image

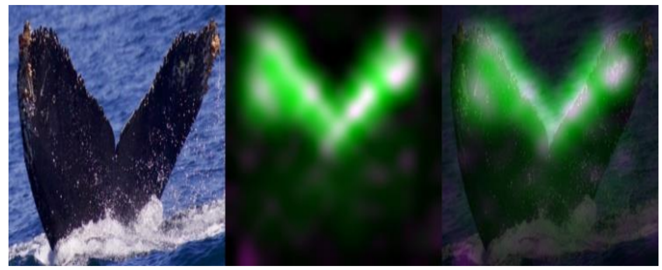
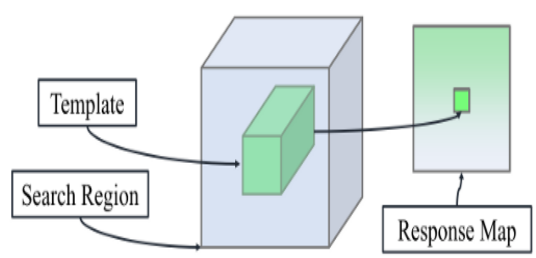
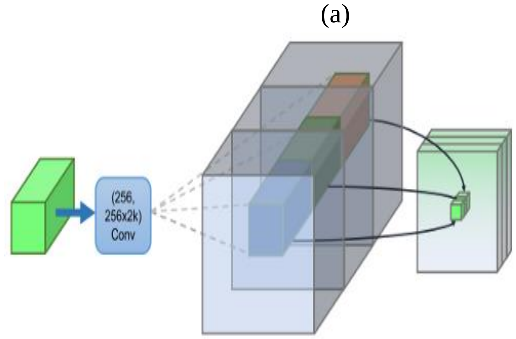


Fig. 6. SiameseRPN++ Visualization in RGB image

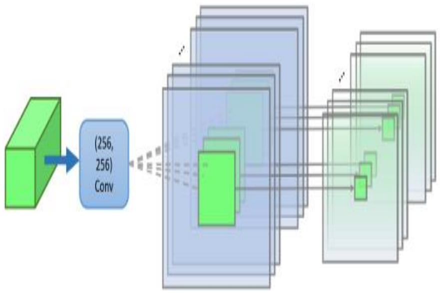
This may also track the object with multiple bounding boxes with various different scores. System produces the special score for an object that is directly reciprocate with 2k logits that exposes k boxes with 2 logits and compute with 4k localization. Here the trained network classifies the object with full proportional with binarization segmentation and Siam network might has been improved with SiamRPN++ and removes the bugs in the network and explores more strategic behavior for analyzing the features for better regression. System pertains the SOTA scores with single object tracking module and system achieved very impressive frame rates upto 35fps as an average perspective and it has been reached to 80 frames per second.



(a)



(b)



(c)

Fig. 7. a,b,c are cross-correlational method in SiamFC, SiamRPN & SiamRPN++ respectively

B. Yolo-v3

YOLO ("You Only Look Once") is a viable constant object acknowledgment calculation, first depicted in the fundamental 2015 paper by Joseph Redmon et al. In this article we present the concept of object identification, the YOLO calculation itself, and one of the calculation's open source executions: Darknet. Image classification is one of the many energizing uses of convolutional neural networks. Beside straightforward image classification, there are a lot of captivating issues in PC vision, with object identification being perhaps the most fascinating. It is commonly connected with self-driving vehicles where frameworks mix PC vision, LIDAR and different advances to produce a multidimensional portrayal of the street with every one of its members. Object discovery is likewise commonly utilized in video reconnaissance, particularly in swarm checking to forestall psychological militant assaults, tally individuals for general insights or investigate client experience with strolling ways inside retail plazas.

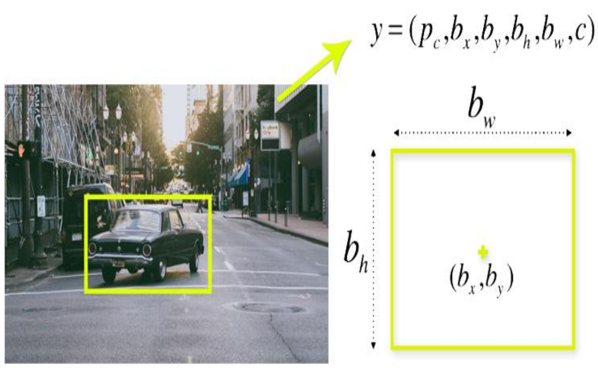


Fig. 8. Yolo Object Detection & Tracking

C. Flow Chart

First of all, system attains a video frame and later pre-processes it for better appearance such as histogram equalization, grayscaling. Once the enhancement has been completed then system loaded the SiameseRPN++ network along with Yolo-v3 for object detection and tracking. Then system will validate the blob with groundtruth for correct tracking. System is reconsolidated with two different techniques i.e. SiamRPN++ and Yolo-v3 for better pattern recognition and object tracking. If bounding box retains the sliding window then it would be considered as correct regression otherwise count as overlap or lost.

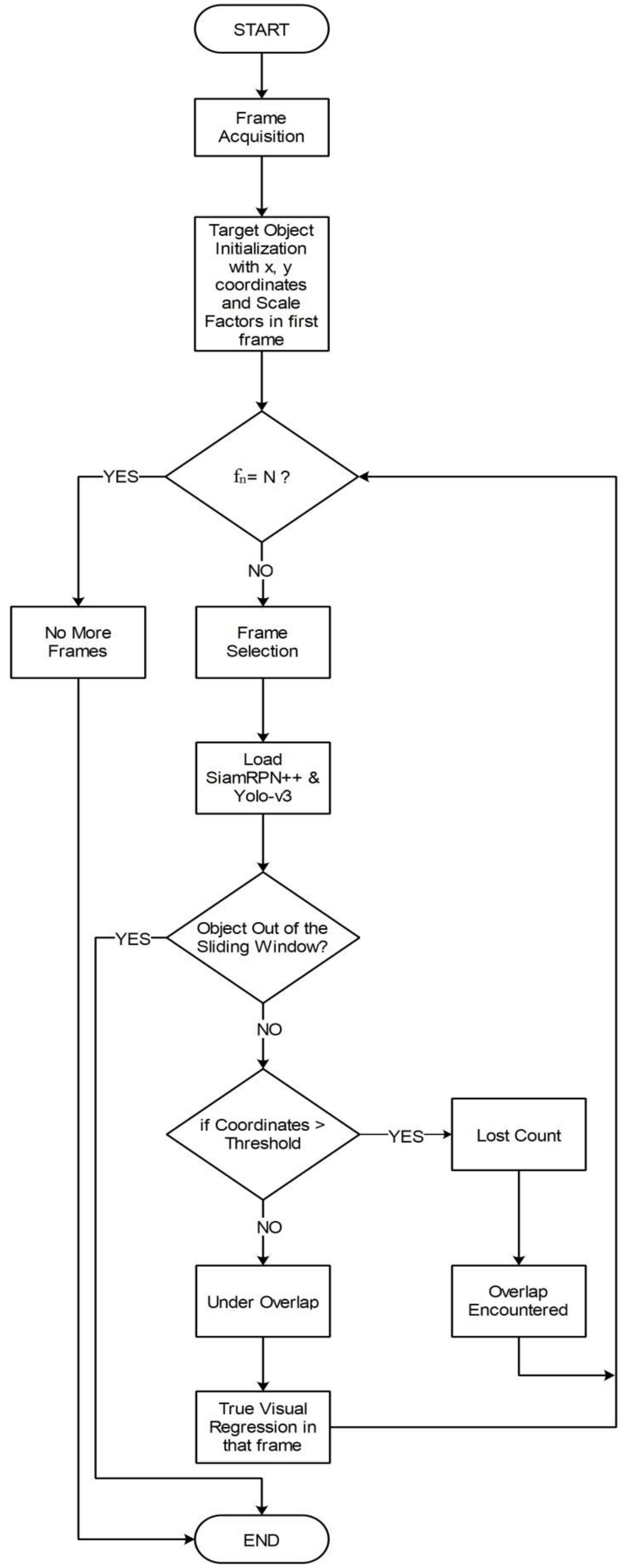


Fig. 9. Flow Chart of Proposed Work

SiameseRPN++ and Yolo-v3 Algorithm
Initialization
Input: 2-D First Frame `
Output: Object Regression till last Frame
Step 1: Input First 2-D Image Matrix
Step 2: Convert RGB image to grayscale image
Step 3: Target Object as per Groundtruth as bounding box b_1
Step 4: Frame sequences of a dataset $\{X_t\}_{t=1}^{T}$, X_1 is the first frame
Step 5: Load Siamese-RPN++ and Yolo-v3 trained network
Step 6: for *f=1 to N* ***do***
Search region x in X_t using pattern recognition;
if x > *Threshold* **then**
count Overlap++;
else
count True Regression++;
end else
end if
Step 7: Extract {*Cx, Cy, X, Y} Co-ordinates*
Where Cx & Cy are x and y co-ordinates, X, Y are the width and height respectively.
Step 9: Compare Extracted Co-ordinates with Groundtruth
Step 10: Compute Accuracy with lost count and true Regression
Step 11: End

V. RESULT ANALYSIS

The system has been tested with TB50 and TB100 benchmarks where 100 of video challenges available with more than 74,000 frames. The performance has been evaluated in the terms of average overlap, loss and accuracy. Each challenge of all benchmarks pertain distinct frame rate.

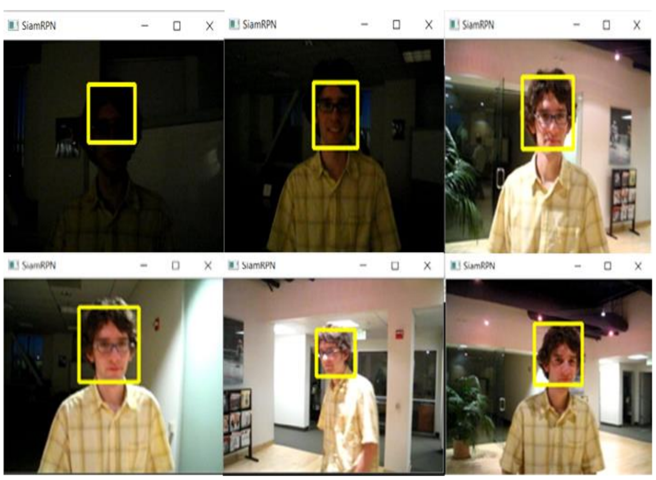


Fig. 10. Proposed Tracking Result for David- TB50 Benchmark

Table No. I shows the performance of benchmark OTB50 and as per all the challenges; the mean accuracy is recorded as 91.17 %. 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. There are bit challenges where system suffers but most of the challenges have been successfully handled by the proposed system with good preciseness and frame execution speed.

Table No. I Recorded Accuracy of Benchmark OTB50

Datasets	***Accuracy***	***Datasets***	***Accuracy***
Basketball	100	Human3	100
Biker	100	Human4	53.37331
Bird1	42.97520661	Human6	92.67677
BlurBody	99.4012	Human9	100
BlurCar2	85.47009	Ironman	75.90361
BlurFace	63.89452	Jump	95.90164
BlurOwl	98.17444	Jumping	100
Bolt	100	Liqour	60.25273
Box	98.27735	Matrix	88
Car1	100	MotorRolling	92.68293
Car4	100	Panda	100
CarDark	100	RedTeam	100
CarScale	96.03175	Shaking	100
ClifBar	96.18644	Singer2	79.5082
Couple	100	Skating1	88.30769
Crowds	100	Skating2	71.67019
David	79.19321	Skiing	100
Deer	100	Soccer	91.46341
Diving	98.13953	Surfur	100
DragonBaby	87.61062	Sylvester	100
Dudek	99.47598	Tiger2	87.12329
Football	54.69613	Trellis	100
Freeman4	93.63958	Walking	100
Girl	100	Walking2	100
		Woman	97.31993
Mean		**91.17040299**	

Table No. II Recorded Overlap/Lost of Benchmark OTB50

Datasets	***Overlap***	***Datasets***	***Overlap***
Basketball	0	Human3	0
Biker	0	Human4	46.62669
Bird1	57.02479	Human6	7.32323
BlurBody	0.5988	Human9	0
BlurCar2	14.52991	Ironman	24.09639
BlurFace	36.10548	Jump	4.09836
BlurOwl	1.82556	Jumping	0
Bolt	0	Liqour	39.74727
Box	1.72265	Matrix	12
Car1	0	MotorRolling	7.31707
Car4	0	Panda	0
CarDark	0	RedTeam	0
CarScale	3.96825	Shaking	0
ClifBar	3.81356	Singer2	20.4918
Couple	0	Skating1	11.69231
Crowds	0	Skating2	28.32981
David	20.80679	Skiing	0
Deer	0	Soccer	8.53659
Diving	1.86047	Surfur	0
DragonBaby	12.38938	Sylvester	0
Dudek	0.52402	Tiger2	12.87671
Football	45.30387	Trellis	0
Freeman4	6.36042	Walking	0
Girl	0	Walking2	0
		Woman	2.68007
Mean		**8.829597008**	

Table No. II shows the overlap or lost of benchmark OTB50 and as per all the challenges; the mean overlap is recorded as 8.82 %.

Table No. III Recorded Accuracy of Benchmark OTB100

Datasets	Accuracy	Datasets	Accuracy
Bird2	100	Freeman1	100
BlurCar1	99.05660377	Freeman3	100
BlurCar3	100	Girl2	45
BlurCar4	92.63157895	Gym	100
Board	59.31232092	Human2	94.41489362
Bolt2	67.91808874	Human5	100
Boy	100	Human7	94.41489362
Car2	100	Human8	100
Car24	100	Jogging	100
Coke	92.78350515	KiteSurf	100
Coupon	40.36697248	Lemming	93.26347305
Crossing	100	Man	100
Dancer	100	Mhyang	100
Dancer2	100	MountainBike	100
David2	100	Rubik	97.39609414
David3	97.61904762	Singer1	100
Dog	84.25925926	Skater	100
Dog1	39.92592593	Skater2	95.40229885
Doll	97.49483471	Subway	100
FaceOcc1	41.59192825	Suv	100
FaceOcc2	76.20650954	Tiger1	99.43502825
Fish	100	Toy	97.78597786
Fleetface	65.62942008	Trans	50
Football1	100	Twinnings	99.78813559
		Vase	87.45387454
Mean			**89.98266663**

Table No. III shows the performance of benchmark OTB100 and as per all the challenges; the mean accuracy is recorded as 89.98 %.

Table No. IV Recorded Overlap/Lost of Benchmark OTB100

Datasets	Overlap	Datasets	Overlap
Bird2	0	Freeman1	0
BlurCar1	0.943396	Freeman3	0
BlurCar3	0	Girl2	55
BlurCar4	7.368421	Gym	0
Board	40.68768	Human2	5.585106
Bolt2	32.08191	Human5	0
Boy	0	Human7	5.585106
Car2	0	Human8	0
Car24	0	Jogging	0
Coke	7.216495	KiteSurf	0
Coupon	59.63303	Lemming	6.736527
Crossing	0	Man	0
Dancer	0	Mhyang	0
Dancer2	0	MountainBike	0
David2	0	Rubik	2.603906
David3	2.380952	Singer1	0
Dog	15.74074	Skater	0
Dog1	60.07407	Skater2	4.597701
Doll	2.505165	Subway	0
FaceOcc1	58.40807	Suv	0
FaceOcc2	23.79349	Tiger1	0.564972
Fish	0	Toy	2.214022
Fleetface	34.37058	Trans	50
Football1	0	Twinnings	0.211864
		Vase	12.54613
Mean			**10.01733337**

Table No. 1 shows the Overlap/Lost of benchmark OTB100 and as per all the challenges; the mean overlap is recorded as 10.01 %. The mean accuracy of TB50 and TB100 are 91.17 % and 89.98 % respectively. The mean overlap of TB50 and TB100 are 8.82 % and 10.01 % respectively. The datasets have 100 videos with more than 70 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.

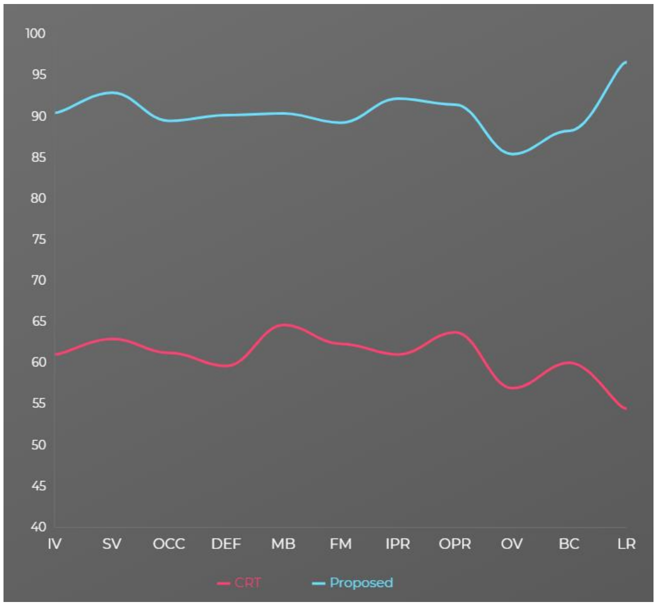
Table No. V Comparison of Evaluations Under 11 Attributes for OTB50: Deformations (DEF), Illumination Variations (IV), Background Clutter (BC), In-Plane Rotations (IPR), Fast Motions (FM), Occlusions (OCC), Out of Plane Rotations (OPR), Motion Blurs (MB), Scale Variations (SV), Out of Views (OV) and Low Resolutions (LR)

TB50 - Mean Accuracy in %		
	CRT	**SiameseRPN++ & Yolo-v3**
IV	61.1	90.51843909
SV	63.0	92.97709189
OCC	61.3	89.53941536
DEF	59.7	90.23027939
MB	64.7	90.4427705
FM	62.4	89.31595179
IPR	61.1	92.25185621
OPR	63.8	91.51780903
OV	57.0	85.49836333
BC	60.1	88.32251474
LR	54.5	96.686389

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

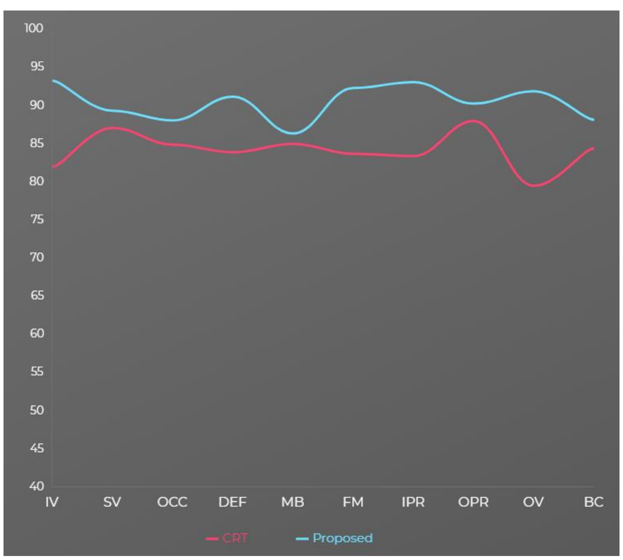
TB100 - Mean Accuracy in %		
	CRT	**SiameseRPN++ & Yolo-v3**
IV	82.0	93.27104318
SV	87.1	89.34040725
OCC	84.9	88.06696191
DEF	83.9	91.19076519
MB	85.0	86.3540672
FM	83.7	92.32306094
IPR	83.4	93.09180635
OPR	88.0	90.27728795
OV	79.5	91.89738508
BC	84.4	88.16666124

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



Graph I 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. II Comparison of Evaluations Under 11 Attributes for OTB100



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

CONCLUSION

In this work, system has been trained and track the desired object using Siamese-RPN++ along with Yolo-v3 for better tracking and pattern recognition. System successfully accepted the challenges proposed in TB50 and TB100 and pertained better level of accuracy as compare to the earlier proposed system i.e. CRT. System scored 91.17% and 89.98% of accuracy for TB50 and TB100 respectively. System has lesser overlap or lost that reaches the better efficiency and frame rate too. 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 Tensorflow or any other object classification or detection technique for better precision in future.

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. Volume: 27, IEEE Signal Processing Magazine. 27 (4): 14–24.

[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.

[3] S. Kang; J. Paik; A. Koschan; B. Abidi & M. A. Abidi (2003). Real-time video tracking using PTZ cameras. Proc. SPIE. Sixth International Conference on Quality Control by Artificial Vision. 5132: 103–111.

[4] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui (2010). Visual object tracking using adaptive correlation filters. In Computer Vision and Pattern Recognition, pages 2544– 2550.

[5] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg (2016). Beyond correlation filters: Learning continuous convolution operators for visual tracking. In European Conference on Computer Vision, pages 472–488.

[6] D. Held, S. Thrun, and S. Savarese (2016). Learning to track at 100 fps with deep regression networks. In European Conference on Computer Vision, pages 749–765.

[7] S. Ren, K. He, R. Girshick, and J. Sun (2015). Faster r-cnn: towards real-time object detection with region proposal networks. In International Conference on Neural Information Processing Systems, pages 91–99.

[8] Manning Free Content Center, The Computer Vision Pipeline, Part 4: feature extraction (2019). Retrieved August 2021, from https://freecontent.manning.com/the-computer-vision-pipeline-part-4-feature-extraction/.

[9] H. Li, S. Wu, S. Huang, K. Lam and X. Xing (2019). Deep Motion-Appearance Convolutions for Robust

Visual Tracking. in IEEE Access, vol. 7, pp. 180451-180466.

[10] Linyu Zheng, Ming Tang, Jinqiao Wang (2018). 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.

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

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

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan (2010). Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645.

[14] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg (2015). Learning spatially regularized correlation filters for visual tracking. in Proc. IEEE Int. Conf. Comput. Vis., pp. 4310–4318.

[15] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer (2014). Adaptive color attributes for real-time visual tracking. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1090–1097.

[16] K. He, X. Zhang, S. Ren, and J. Sun (2015). Deep residual learning for image recognition. CoRR, vol. abs/1512.03385, pp. 1–12.

[17] B. Babenko, M.-H. Yang, and S. Belongie (2011). Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 8, pp. 1619–1632.

[18] H. Grabner, M. Grabner, and H. Bischof (2006). Real-time tracking via on-line boosting. in Proc. Brit. Mach. Vis. Conf., pp. 1–10.

[19] K. Simonyan and A. Zisserman (2015). Very deep convolutional networks for large-scale image recognition. in Proc. Int. Conf. Learn. Represent., pp. 1–14.

[20] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista (2015). High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583–596.

[21] S. Hare, A. Saffari, and P. H. S. Torr (2011). Struck: Structured output tracking with kernels. in Proc. IEEE Int. Conf. Comput. Vis., pp. 263–270.

[22] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista (2012). Exploiting the circulant structure of tracking-by-detection with kernels. in Proc. Eur. Conf. Comput. Vis., pp. 702–715.

***