Deep Reinforcement Learning for Action Based Object Tracking in Video Sequences

***Aishwarya N Inamdar*[1]*, Vindhya M .P*[2]**
[1]*PG Student,* [2]*Associate Professor*
[1,2]*Department of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India*
Email: *aishwaryainamdar918@gmail.com*

DOI: *http://doi.org/10.5281/zenodo.3251617*

Abstract

In this paper, we propose a valuable route for visual object tracker which catches a bounding box to zone of premium physically in the video frames by recognizing the activity got the hang of utilizing the convolution neural systems. The proposed convolution neural network used to control tracking actions is done with various training video sequences and fine-tuned during the actual tracking of the object. Pretrain of the video is done using deep reinforcement learning (RL) along with the supervised learning. Mostly named information from the RL can be utilized for semi supervised learning and assessing through object tracking benchmark dataset, the proposed tracker is confirmed to accomplish a good performance. The proposed method, which operates in real time on without graphics processing unit, outperforms the state of real time trackers with proper accuracy with performance 10%.

Keywords: *Convolution neural, network, reinforcement learning, supervised learning, Semi supervised learning*

INTRODUCTION

The aim of the visual object tracking is to find a bounding box containing the object of interest is moving along with the bounding box in video, which is one of the main problems in the field of computer vision filed. In these ongoing years, there have been numerous progressions in object tracking algorithm, yet there exist many testing issues amid the various following impediments, for example, movement Blur, impediment, enlightenment change, and background mess. Particularly, CNN method the above mention obstacles are caused because of their insufficient feature representations.

The following strategies which use the convolution neural systems (CNNs) which have been utilized to propose robust tracking and a large portion of the improvement in execution alongside utilization of highlight portrayal by profound shrouded layers of the CNN. The initial work uses pretrained CNNs, which are usually trained on a huge scale classification of the dataset [1]. We have proposed a method that precisely selects color features that best discriminate object from the current background and foreground. Utilizing on-line adjustment, we face floating as one of the key issues. Each time we make an update to our tracker a mistake may be presented, bringing about a failure in the object tracking, which may gather after some time bringing about following failure [2].The estimation of the motion trajectory for a target in subsequent video frames of the video sequence, when the state of the target object is given only in the first frame image of the video sequence. Compared to the static object recognition and detection in an image, temporal component in videos provides an important clue for the recognition, detection, and tracking of a moving object in an image sequence. In this way, we can accomplish various valuable data from the moving direction related with the fleeting segment acquired from the image frames. The test results show that it is increasingly alluring to perform spatial element and transient component undertakings mutually to enable them to benefit by one another [7]. Hence, in our work, given an input video sequence, we directly extract the spatial feature from a single image and temporal feature from a serial of images constituted by successive video frames before combing with CF.

One regular issue in surveying tracking calculations is that the revealed outcomes are frequently founded on a couple of arrangements with various initialization or parameters. Inaccurate localization of the target occurs frequently as an object detector may be used for locating the object in the first frame. In addition, an object detector may be used to recover from tracking failures by re-initializing the tracker [8]. The communication between the transient part and the detector permits long term tracking even through durable impediments.

RELATED WORK

Visual Object Tracking

As surveyed in and various trackers have shown their performance and effectiveness on various tracking benchmarks. The approach based on tracking-by detection aims to build a discriminative classifier that distinguishes the target from the surrounding background. Typically these methods capture the target position by detecting the most matching position using the classifier.

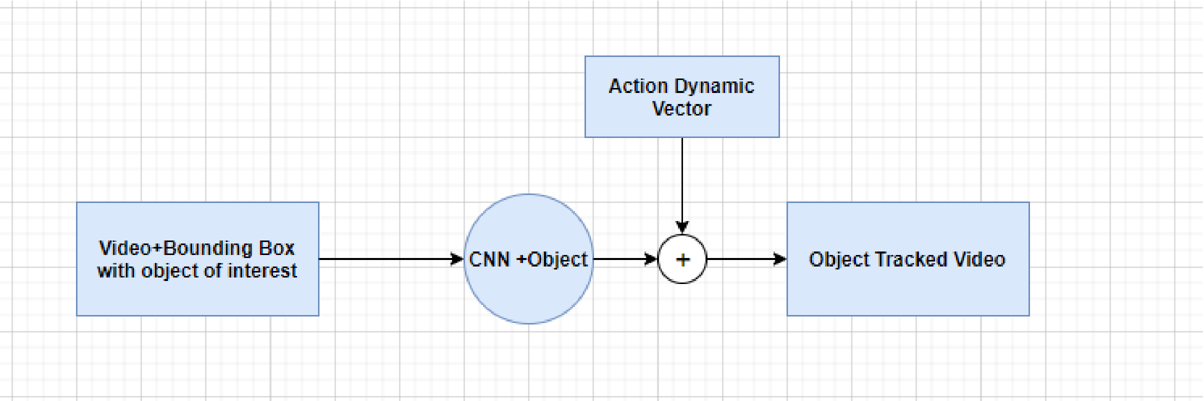

Figure 1: *Basic architecture of the visual object tracking.*

A context-level or level 0 data flow diagram shows the interaction between the system and external agents which act as data sources and data sinks. Basic requirement for project is the input, system and the output device. Here the input can be given in the form of video and feed as the input to the system which consist of the CNN and there the processing of the programs will take place in the system and finallythe output is obtained which is in the form of the video where the object of interest is tracked. After intersection over union calculation object marked with bounding box is tracked. For each iteration action dynamic vector history is created and updated and that history of vector is informed to the action dynamic vector which is already stored outside CNN, this iteration of object tracking is continued till the end of the video frame.

Deep Reinforcement Learning

The goal of RL is to learn a policy that decides sequential actions by maximizing the cumulative future rewards. A Recent trend in RL field is to utilize the deep neural networks as a function approximationof policy or value function. By resorting to the use of deep features, many difficult problems, such as playing Atari games or Go, can be successfully solved in a SS setting.The profound RL methodology has been connected to other computer vision applications, for example, object confinement and activity acknowledgment. In visual tracking field, Zhang et al.have endeavored to use a profound RL approach for trackng objects in video successions. This method focuses on localizing the target by a regression frame by frame; therefore, it does not consider sequential dynamics of the target objects and cannot utilize the weekly SL strategy, which are distinctive aspects in our method[1].A number of tracking methods based on color histograms have been developed. Comaniciu et al. applied the mean shift algorithm to object tracking on the basis of a color histogram. Collins extended the mean shift tracking algorithm to deal with the scale variation of target objects.The latest generation of Convolutional Neural Networks (CNN) have achieved impressive results in challenging benchmarks on image recognition and object detection, significantly raising the interest of the community in these methods. By and by, it is as yet indistinct how extraordinary CNN techniques contrast and one another and with past cutting edge shallow portrayals, for example, the Bag-of-Visual-Words and the Improved Fisher Vector. This paper leads a thorough assessment of these new procedures, investigating diverse profound models and looking at them on a shared conviction, recognizing and uncovering significant execution subtleties. We distinguish a few valuable properties of CNN-based portrayals, including the way that the dimensional of the CNN yield layer can be decreased altogether without adverse affecting execution time of the procedures.We furthermore perceive portions of significant and shallow procedures that can be adequately shared. In particular, we exhibit that the data development frameworks typically associated with CNN-based procedures can in like manner be associated with shallow systems, and result in a similar to execution help. Source code and models to repeat the preliminaries in the paper is made unreservedly open. The objective of moving object area is to take a video course of action from a fixed/moving camera and yields a twofold spread addressing moving object for each edge of the progression. In any case, this is definitely not a basic endeavor to do as a result of various challenges and inconveniences included when a camera is utilized to get a video course of action of moving object. Here, we present these troublesome issues in nuances.

Proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial P number of frames are used for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are the outputs of this filter from which the velocity and acceleration granules are generated. The object location in the next frame is estimated based on these information and the roughness in estimation is checked. If it is high then the intutionistic entropy is computed for the boundary granules and the ones which are found ambiguous are eliminated in order to track the object in the frame accurately. The tracking results, thus produced, are comparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusion. The proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are the outputs of this filter from which the velocity and acceleration granules are generated. The object location in the next frameis estimated based on this information and the roughness in estimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which are found ambiguous are eliminated in order to track the object in the frame accurately. The tracking results, thus produced, are comparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusion. The proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are the outputs of this filter from which the velocity and acceleration granules are generated. The object location in the next frameis estimated based on this information and the roughness in estimation is checked. If it is high then the intutionistic entropy is computed for the boundary granules and the ones which are found ambiguous are eliminated in order to track the object in the frame accurately. The tracking results, thus produced, are comparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusion. The proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are the outputs of this filter from which the velocity and acceleration granules are generated. The object location in the next frame is estimated based on this information and the roughness in estimation is checked. If it is high then the intutionistic entropy is computed for the boundary granules and the ones which are found ambiguous are eliminated in order to track the object in the frame accurately. The tracking results, thus produced, are comparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusion. The proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames is used for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are the outputs of this filter from which the velocity and acceleration granules are generated. The object location in the next frame is estimated based on this information and the roughness in estimation is checked. If it is high then the intutionistic entropy is computed for the boundary granules and the ones which are found ambiguous are eliminated in order to track the object in the frame accurately. The tracking results, thus produced, are comparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusion. The proposed method works as follows. Given an in-put video sequence, spatio-colorn eighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are the outputs of this filter from which the velocity and acceleration granules are generated. The object location in the next frame is estimated based on this information and the roughness in estimation is checked. If it is high then the intutionistic entropy is computed for the boundary granules and the ones which are found ambiguous are eliminated in order to track the object in the frame accurately. The tracking results, thus produced, are comparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are the outputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frame is estimated based on this information and the roughness in estimation is checked. If it is high then the intutionistic entropy is computed for the boundary granules and the ones which are found ambiguous are eliminated in order to track the object in the frame accurately. The tracking results, thus produced, are comparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial P number of frames are used for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial P number of frames are used for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionthe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. It’s initial P number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusion

PROPOSED SYSTEM

Visual Based Object Tracking

Proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial P number of frames are used for initial object labelling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusion.The proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-colorn eighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter.The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which are found ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. It’s initial P number of frames are used for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on theseinformation and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusionThe proposed method works as follows. Given an in-put video sequence, spatio-color neighborhood granules are formed over all the frames. Its initial number of frames areused for initial object labeling with the proposed NRS filter. The color model and velocity profiles of the objects are theoutputs of this filter from which the velocity and accelerationgranules are generated. The object location in the next frameis estimated based on this information and the roughness inestimation is checked. If it is high then the intutionistic entropyis computed for the boundary granules and the ones which arefound ambiguous are eliminated in order to track the object inthe frame accurately. The tracking results, thus produced, arecomparable to those of some of the state-of-the-art algorithms, while superior in case of total occlusio. Proposed WorkReinforcement Lea

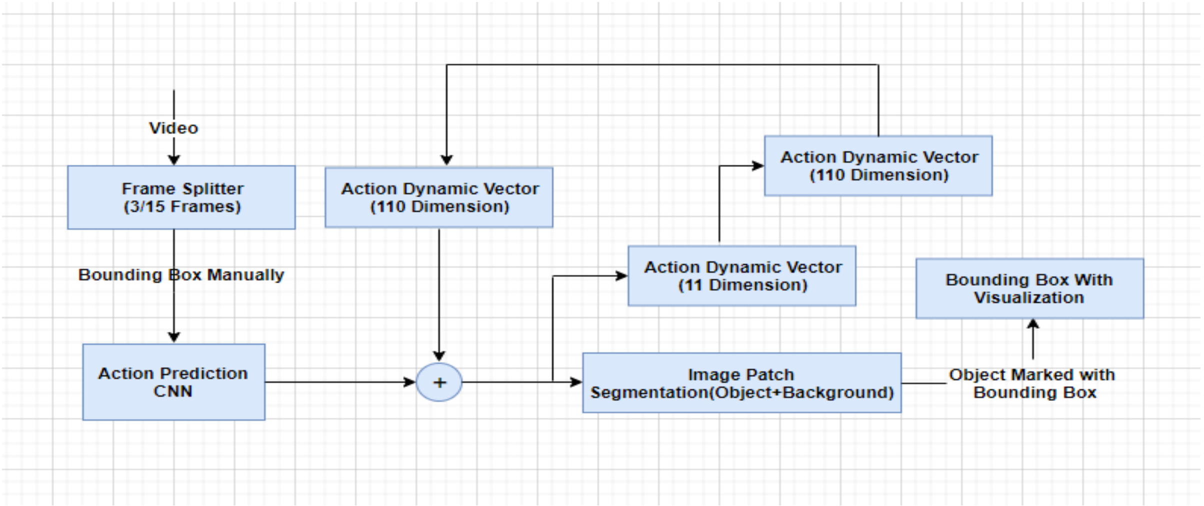

***Figure 2:** Architecture of deep reinforcement learning for action based object tracking with video sequence.*

Fig. 2 describes the detailed architecture of the proposed system .The basic architecture needs the input devices, system and the output devices.We are giving the video as the input and that video is converted in to number of image frames F, to CNN the 3 to 5 frames are sent.CNN uses the 112*112*3 as the initial buffer of the frame and is converted in to one single vector. The bounding box is done manually and then processing of the video will take place. We are using the dynamic vector where the action will be stored there will be 110 dimension.At the end of each iteration of CNN output is concatenated with the Action dynamic vector Dt After the concatenation the system will produce the two vectors where one with 11 dimension action dynamic vector and other with the 2 dimension vector which will depict the image patch with the bounding box this will be considered as the foreground of the image frame and excluding the bounding box of the image frame is taken as the background. This can be calculated using the intersection over union.

$$r(sT) = \begin{cases} 1, & if\ IoU(bT, G \\ -1, & oth \end{cases} 1$$

EXPERIMENTS AND RESULTS

The tracking of the object begins with the second frame of the video that tracking will continued till the last frame of the video. Once the last frame of the video is encountered the tracking of the visual based tracking with deep enforcement will stop. The final output will be in the form of the video.

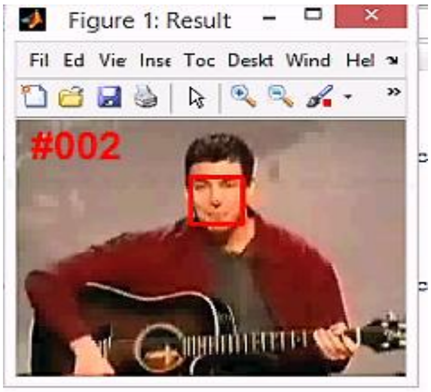

***Figure 3:** Object tracking start from second.*

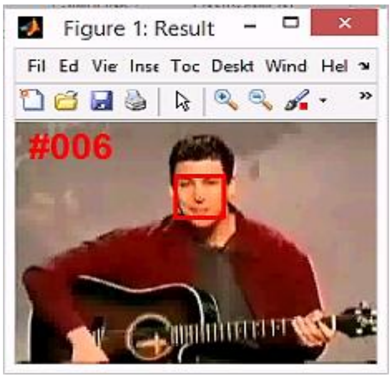

***Figure 4:** Example of the success frames.*

Self-Evaluation

To verify the effectiveness of the components of ADNet, we conducted four variants of ADNet and evaluated. In ADNet-init, the parameters of convolutional networks (conv1–conv3) are initialized with the VGG-M model, and the fully connected layers (fc4–fc7) are initialized with random noises. "ADNet + SL" is the pretrained models with SL using fully labeled frames of the training sequences.To examine the tracking performance according to the type of the reward function, we conducted an "ADNet-R (cont)," whose reward function was defined as a continuous function. The continuous reward function uses IoU scores as follows:

r(sT)=IoU(bT,G) 2

The following two weakly correlated performance measures are used due to their high level of interpretability (i) accuracy and (ii) robustness. The accuracy of the system will measures how well bounding box predicted by the tracker overlaps with the already defined values of bounding box. The vigor of the framework can be characterized as how frequently tracker loses the object of interest (fails) amid following. Failure of object tracking can be distinguished when overlapping of the bounding box measure become zero. Because of the disappointment in the following the tracker gets introduced past last 5 edge of the tracker.

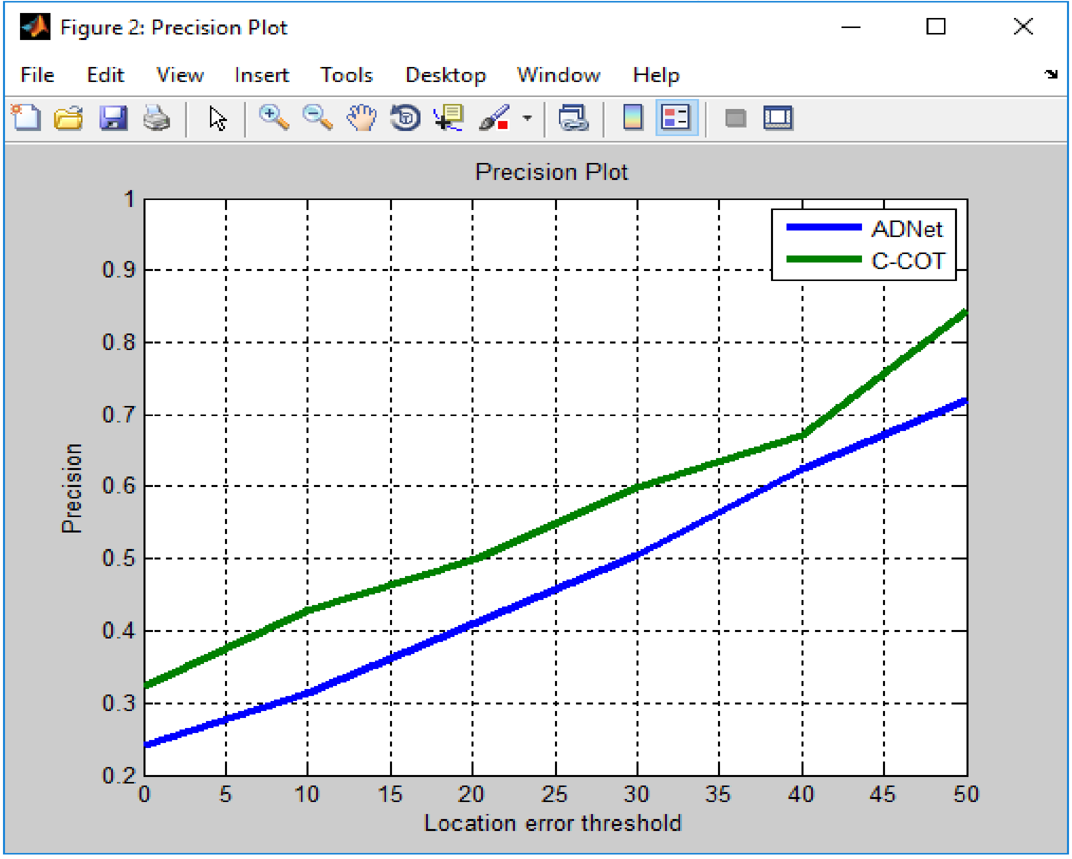

Figure 5: *Precision OPE Graph.*

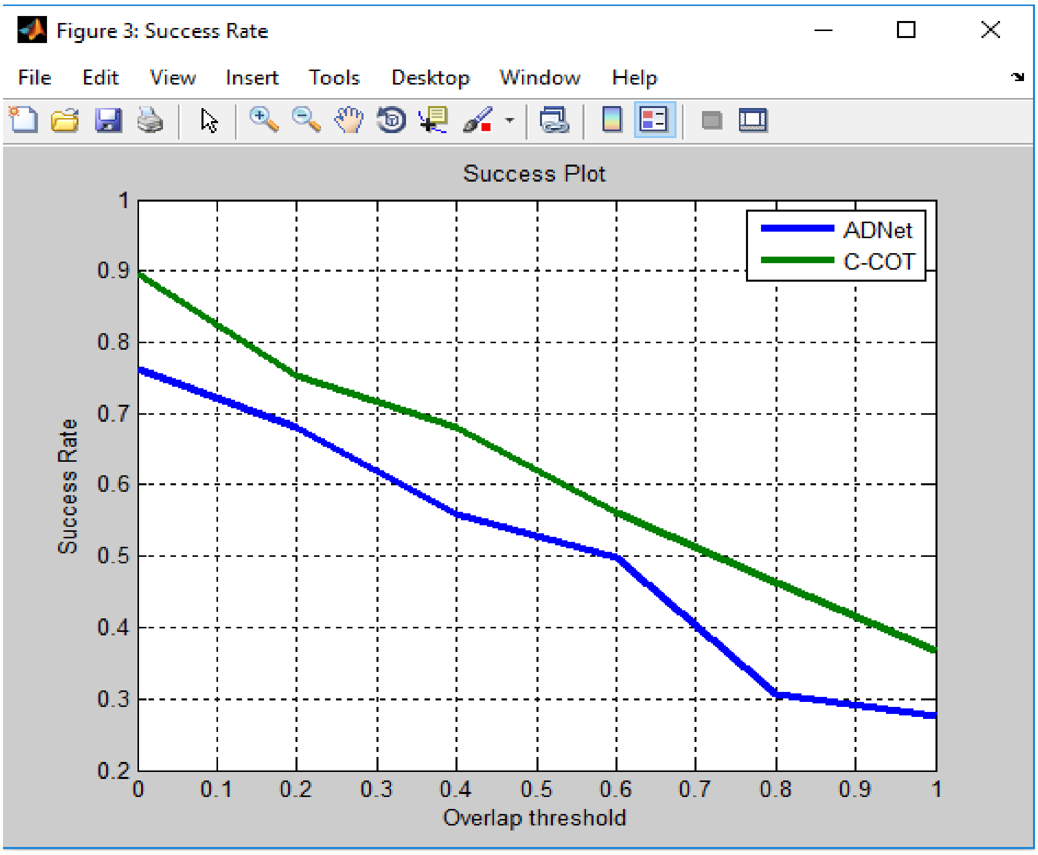

***Figure 6:** Success OPE graph.*

CONCLUSION

The action-driven tracking strategy makes a significant contribution to the reduction of computation complexity in tracking. In addition, RL makes it possible to use partially labeled data, which could greatly contribute to the building of training data with a little effort. According to the evaluation results, the proposed tracker achieves the state-of-the-art performance.

REFERENCES

1. Sangdoo Yun, Jongwon Choi,Youngjoon Yoo, Kimin Yun, Jin Young Choim (June 2018), “Action-Driven Visual Object Tracking With Deep Reinforcement Learning”,*IEEE transactions on neural networks and learning systems*, Volume 29, Issue 6.
2. H. Grabner, M. Grabner, H. Bischof (2006), “Real-time tracking via on-lineboosting”,*Brit. Mach. Vis. Conf.*,Volume 1, Issue 5, pp. 6.
3. B. Babenko, M.-H. Yang, S. Belongie (Aug. 2011), “Robust object tracking withonline multiple instance learning ”, *IEEE Trans. Pattern Anal. Mach. Intell.*,Volume 33, Issue 8, pp. 1619–1632.
4. Z. Kalal, K. Mikolajczyk, J. Matas (July 2012), “Tracking-learning-detection”, *IEEE Trans. Pattern Anal. Mach. Intell.*,Volume 34, no. 7, pp. 1409–1422.
5. D. S. Bolme, J. R. Beveridge, B. A. Draper, Y. M. Lui, “Visualobject tracking using adaptive correlation filters”,*IEEE Conf.*
6. Y. Wu, J. Lim, M. H. Yang (Sep. 2015), “Object tracking benchmark”,*IEEETrans. Pattern Anal. Mach. Intell.*,Volume 37, Issue 9, pp. 1834–1848, *Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2544–2550.
7. J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, J. Y. Choi (Jul. 2017), “Attentional correlation filter network for adaptive visual

tracking”,*IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4828–4837.

8. Y. Wu, J. Lim, M. H. Yang (Sep. 2015), “Object tracking benchmark”,*IEEE Trans. Pattern Anal. Mach. Intell.*,Volume 37, Issue 9, pp. 1834–1848.
9. M. Kristan, et al. (Mar. 2014), “The visual object tracking VOT2014 challenge results”,*Eur. Conf. Comput. Vis. Workshops*, Berlin, Germany, pp. 191–217.
10. K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman. (2014), “Return of the devil in the details: Delving deep into convolutional nets”, [Online]. Available: https://arxiv.org/abs/1405.3531.

Cite this article as: *Aishwarya N Inamdar, & Vindhya M .P. (2019). Deep Reinforcement Learning for Action Based Object Tracking in Video Sequences. Journal of Web Development and Web Designing, 4(2), 20–30. http://doi.org/10.5281/zenodo.3251617*