Visual Object Tracking in Drone Images with Deep Reinforcement Learning

Derya GÖZEN and Sedat OZER
Bilkent University

- Drones are in demand for use in many domains from agriculture to surveillance.
- Drone images introduce additional challenges in object tracking such as smaller object sizes, different orientations and angle of views, which makes it difficult for existing object trackers to be directly applied to drone images.
- The recent focus for object tracking is on developing reinforcement learning (RL) based solutions.
- We propose a deep action sequence based RL tracker for drone images.

Motivation

- Growing number of drone applications
- New challenges with tracking in drone images
- Limited number of papers on tracking in drone images
- Potential of RL-based solutions on learning meaningful actions

Tracking with RL

The action-decision network, ADNet, determines the future position of the object of interest in terms of a predicted action sequence, and each action in the sequence is predicted from the current state (Fig. 1). ADNet is trained by both supervised and RL techniques, and an online adaptation process in tracking is adopted during the testing process.

1 Supervised learning stage {w1, w2,..., w7}
2 Reinforcement learning stage {w1, w2,..., w6}
3 Online adaptation in tracking {w4, w5, w6, w7}

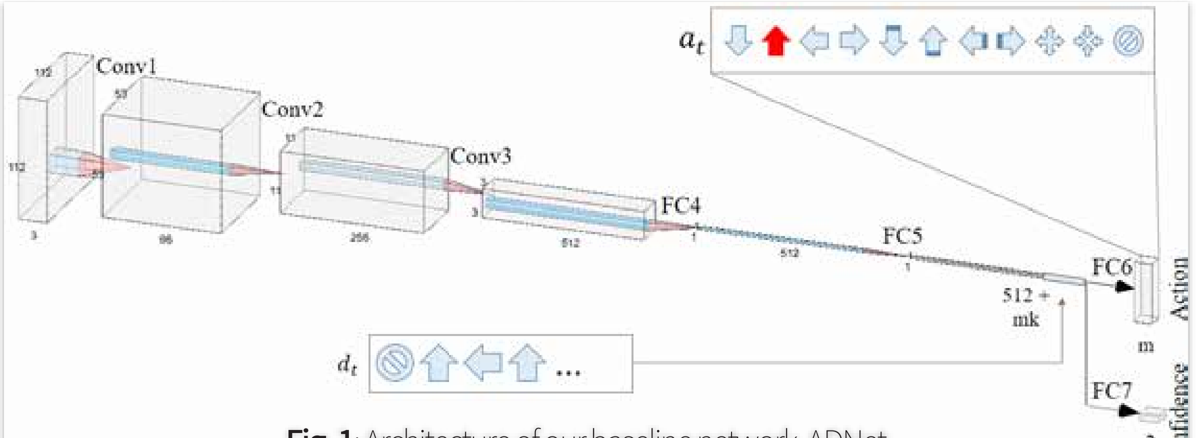

Fig. 1: Architecture of our baseline network, ADNet.

ACKNOWLEDGMENT

This paper has been produced benefiting from the 2232 International Fellowship for Outstanding Researchers Program of TÜBİTAK (Project No: 118C356). However, the entire responsibility of the paper belongs to the owner of the paper. The financial support received from TÜBİTAK does not mean that the content of the publication is approved in a scientific sense by TÜBİTAK.

Our Action-Sequence-Based RL Trackers

We introduce four different action-sequence-based RL trackers.

Model A

Our first model focuses on utilizing and choosing proper action sets for the drone domain (Fig. 2).

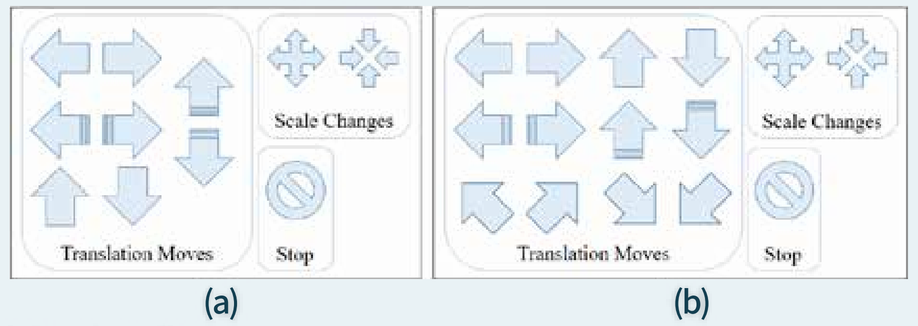

Fig. 2: The set of actions defined for (a) our base-network, and for (b) Model-A.

Model B

ADNet architecture uses the first three convolutional layers of VGG-M for initialization whereas Model-B, uses VGG-F as backbone network (Fig. 3).

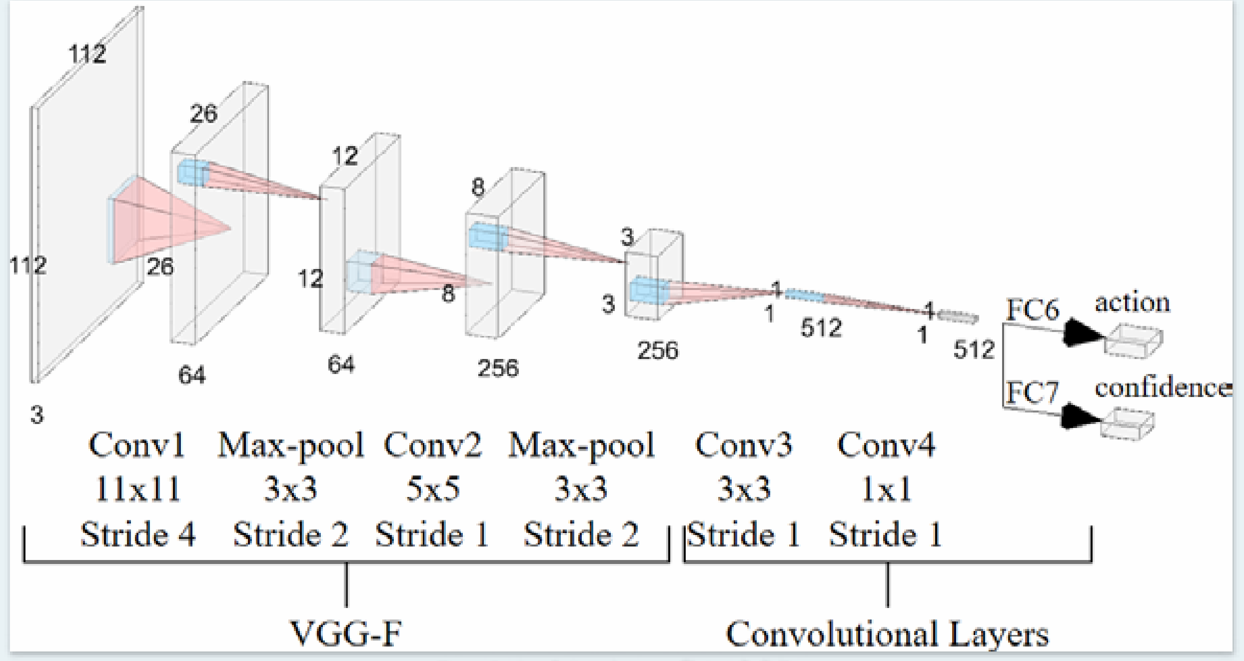

Fig. 3: Architecture of Model-B.

Model C

Firstly, the reward function of Model-C is defined as following:

$$r(s_T) = \begin{cases} (10 - length\{a_{t,l}\}) * IoU(b_t, G), & if\ IoU(b_t, G) > 0.7 \\ -1, & otherwise \end{cases}$$

Secondly, we let reinforcement learning algorithm give reward to the set of actions each time the tracking is applied on a sample video clip to update the model.

Model D

Model-D is very similar to Model-C in terms of reinforcement learning algorithm, but in Model-D also updates. However, its reward function is identical to that of our base-network, ADNet:

$$r(s_T) = \begin{cases} +1, if\ IoU(b_t, G) > 0.7 \\ -1, otherwise \end{cases}$$

Fig. 4: Sample successful cases of our Model-C over the baseline network (ADNet) on VisDrone2019 data set. The blue, red and green bounding boxes represent the bounding boxes of ADNet, Model-C and the ground truths, respectively.

Data Sets

Our RL-based deep trackers are;

1. trained on 58 videos collected from VOT2013, VOT2014 and VOT2015 data sets, and tested on OTB, which includes a total number of 100 videos from both OTB-50 and OTB-100 data sets.
2. trained on VisDrone2019 (VisDrone2019 SOT trainset part1) data set, which includes 43 aerial videos for training, and tested on 11 aerial videos (from VisDrone2019-SOT valset).

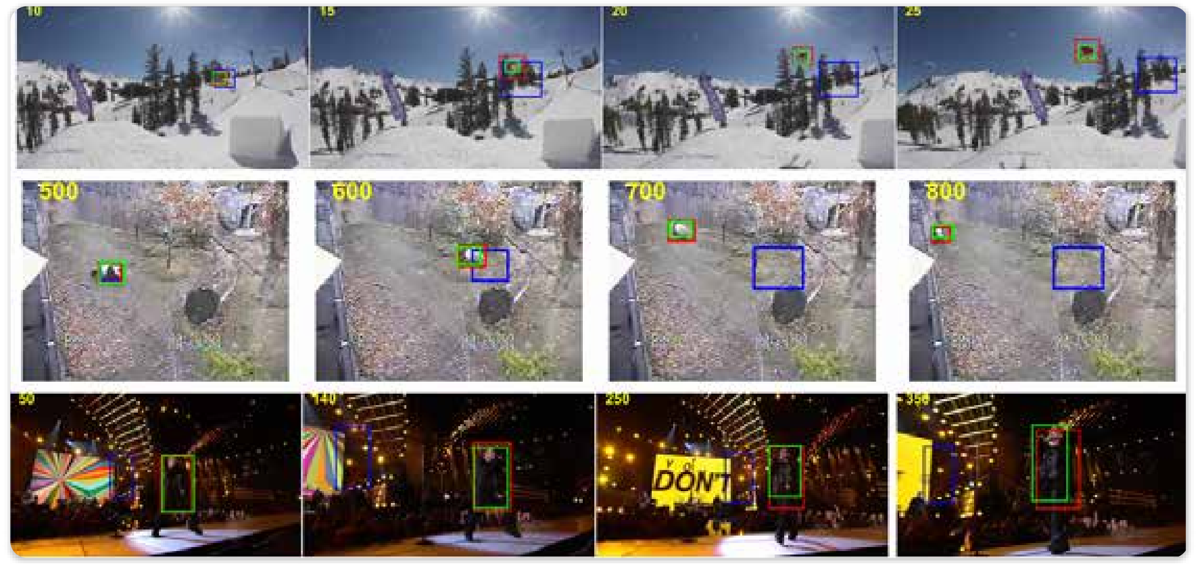

Fig. 5: Sample successful cases of our Model-D over the baseline network (ADNet) on OTB-100 data set. The blue, red and green bounding boxes represent the bounding boxes of ADNet, Model-D and the ground truths, respectively.

Results & Analysis

Overall Performance

TABLE I: Comparison of our proposed methods to the baseline algorithm on OTB-100 and VisDrone2019 data sets.

Experiment Type	Model	OTB100 Precision (20 pixels)	OTB100 FPS	OTB100 IoU	VisDrone2019 Precision (20 pixels)	VisDrone2019 FPS	VisDrone2019 IoU
Baseline model	ADNet	78.47%	4.89	0.603	89.15%	6.33	0.579
Action set	Model-A	79.45%	4.58	0.612	91.94%	6.08	0.557
Backbone networ	Model-B	77.15%	8.11	0.574	89.67%	6.53	0.553
Reward function	Model-C	80.61%	6.25	0.589	93.02%	5.61	0.611
	Model-D	81.62%	7.02	0.616	91.74%	6.13	0.615

Challenging Aspects

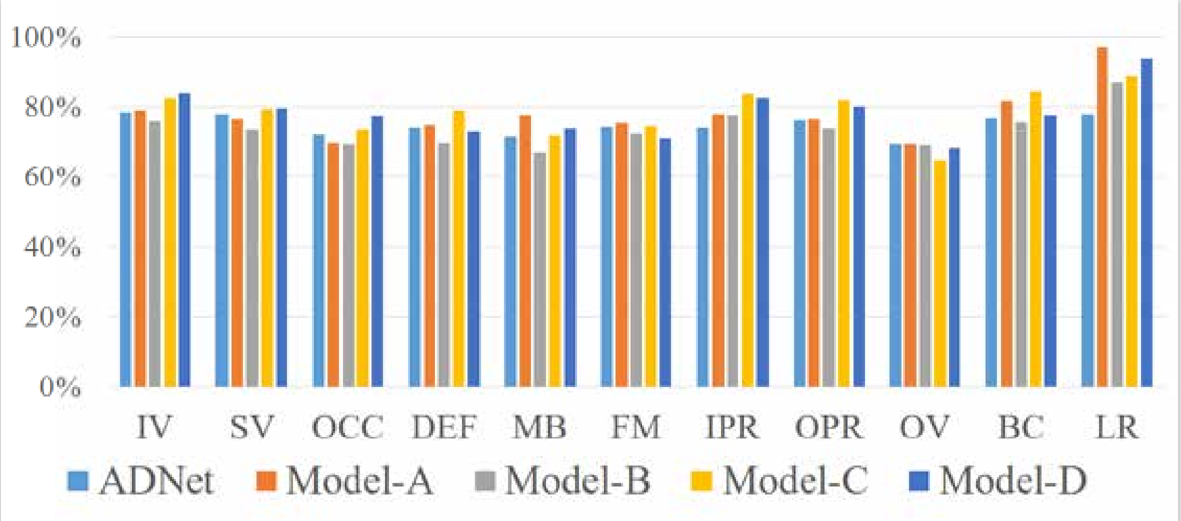

Fig. 6: Average precision results of ADNet, Model-A, Model-B, Model-C, and Model-D across the set of videos from OTB-100, grouped by challenging aspects.