MULTIPLE VEHICLES AND PEOPLE TRACKING IN AERIAL IMAGERY USING STACK OF MICRO SINGLE-OBJECT-TRACKING CNNS

R. Bahmanyar, S. M. Azimi, P. Reinartz

Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany
(Reza.Bahmanyar, Seyedmajid.Azimi, Peter.Reinartz)@dlr.de

KEY WORDS: Aerial Imagery, Vehicle Tracking, Person Tracking, CNNs, Disaster Management, Traffic Management

ABSTRACT:

Geo-referenced real-time vehicle and person tracking in aerial imagery has a variety of applications such as traffic and large-scale event monitoring, disaster management, and also for input into predictive traffic and crowd models. However, object tracking in aerial imagery is still an unsolved challenging problem due to the tiny size of the objects as well as different scales and the limited temporal resolution of geo-referenced datasets. In this work, we propose a new approach based on Convolutional Neural Networks (CNNs) to track multiple vehicles and people in aerial image sequences. As the large number of objects in aerial images can exponentially increase the processing demands in multiple object tracking scenarios, the proposed approach utilizes the stack of micro CNNs, where each micro CNN is responsible for a single-object tracking task. We call our approach Stack of Micro-Single-Object-Tracking CNNs (SMSOT-CNN). More precisely, using a two-stream CNN, we extract a set of features from two consecutive frames for each object, with the given location of the object in the previous frame. Then, we assign each MSOT-CNN the extracted features of each object to predict the object location in the current frame. We train and validate the proposed approach on the vehicle and person sets of the KIT AIS dataset of object tracking in aerial image sequences. Results indicate the accurate and time-efficient tracking of multiple vehicles and people by the proposed approach.

1. INTRODUCTION

Multi-person and -vehicle tracking has several applications such as large-scale event and traffic monitoring, disaster management, and predictive traffic and crowd modeling. Tracking of all vehicles and people in aerial imagery can provide valuable information about the traffic and crowd situation on the ground as aerial imagery allows capturing images of large areas in a very short time. Object tracking in visual data, locating objects of interest in sequences of video frames, is called Visual Object Tracking (VOT). VOT has attracted many research works for a long time and it is still an unsolved problem due to the existing challenges such as cluttered background as well as considerable variations in viewpoints, illuminations, and occlusion.

VOT methods can be categorized into single- and multiple-object tracking (SOT and MOT) methods. While in SOT, we track a single object throughout a given image sequence, in MOT, we track multiple objects within the image sequence at the same time. In ground imagery, for instance in the autonomous driving, MOT is a key element for an autonomous vehicle to plan its path by tracking the trajectories of dynamic objects such as people and vehicles. In recent years, the VOT methods based on deep learning specifically Convolutional Neural Networks (CNNs) (Girshick et al., 2014, Girshick, 2015, Ren et al., 2015, Lin et al., 2017) have shown promising performances in MOT scenarios (Wojke et al., 2017, Bewley et al., 2016). However, most of these methods suffer from high computational costs and slow processing, especially extracting features from each candidate object locations in every frame (El-Shafie et al., 2019). The complexity increases exponentially by increasing the number of objects. In order to employ CNNs for VOT purposes, one approach is to train CNNs as object versus background classifiers in an online manner and apply them to a number of sampled candidate regions, where the region with the highest classification score is then selected as the most visually similar region (Nam, Han, 2016). As the classifier is trained online on the object of interest, this approach could be slow depending on the number of objects and the CNNs' complexities. The offline-training-based approaches however, are much faster than the online-training-based ones. In these approaches, CNNs are trained to regress to the new object positions by getting the cropped area of the previous frame centering at the object and the crop of the current frame as two inputs (Held et al., 2016). The outputs of the CNNs are the bounding box positions of the objects in the current frame.

Figure 1. Sample frames from the vehicle and person tracking sets of the KIT AIS dataset

The tracking approaches are also categorized into short- and long-term tracking. In short-term tracking according to the definition provided by (Kristan et al., 2016), there is not re-detection module meaning that the object is always present throughout the sequence. In contrast, in the long-term tracking, partial occlusion or disappearance does not stop the tracking method (Benfold, Reid, 2011). In this category, methods can re-detect the object after reappearance. The object detectors, including CNN-based ones, could be used to re-detect the objects after their

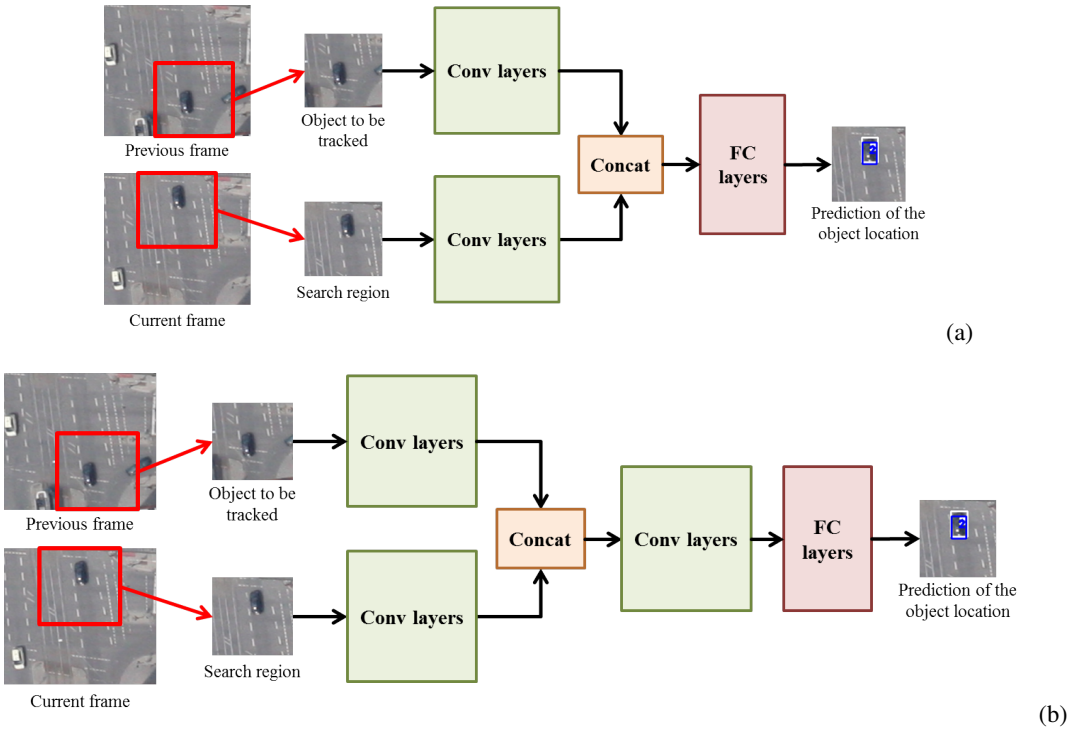


Figure 2. Overview of (a) GOTURN and (b) our proposed modification to GOTURN.

reappearance. For instance, GOTURN (Held et al., 2016) and MDNet (Nam, Han, 2016) are short-term and KCF (Henriques et al., 2014) is a long-term tracker. In addition, trackers can be categorized into offline and online trackers (Luo et al., 2014). While the online trackers only consider the current and previous frames, the offline ones make use of all frames including the future ones.

In this work, we address geo-referenced real-time tracking of multiple vehicles and people in aerial imagery. Specifically speaking, we tackle the task of MOT in a short-term scenario as our method does not include re-detection module. Object tracking in aerial imagery introduces additional challenges such as small size of the objects and the limited temporal resolution of geo-referenced datasets. Furthermore, due to their wider field of view, the number of objects in aerial imagery is usually large which can exponentially increase the processing demands especially in the tracking of multiple objects. Taking all into account, the aim of this work is to introduce an approach for speeding up multi-object tracking while takes advantage of the state-of-the-art performance of CNNs. Our approach utilizes the stack of micro CNNs, where each micro CNN is responsible for the tracking of a single object which we call it Stack of Micro Single-Object-Tracking CNNs (SMSOT-CNNs). The stacking mechanism not only allows the complexity to grow linearly as the number of objects increases, but also facilitates parallelism of the process. In our experiments, we focus on the tracking of vehicles and people in various motion and crowd density scenarios. As a base CNN for the SMSOT-CNNs, we modify the GOTURN single-object tracking network (Held et al., 2016). Figure 2 (b) represents the overview of the modified GOTURN.

In ground imagery, the existence of large-scale and diverse tracking datasets such as the MOT17 dataset (Milan et al., 2016) with 33,705 frames allows developing various methods based on deep learning. However, in the aerial imagery domain, the lack of large and diverse tracking datasets limited the development of well-performing object tracking methods. In this work, we use the KIT AIS dataset[1] in our experiments. This dataset comprises a vehicle tracking set with 229 frames and a person tracking set with 190 frames where their frame rates vary around 2 Hz. Figure 1 shows example frames of the KIT AIS dataset. The images are provided by the German Aerospace Center (DLR) using the 3K camera system composed of three standard DSLR cameras (a nadir-looking and two side-looking cameras) mounted on an airborne platform. Due to the different flight altitudes and the camera configurations, the images are with different ground sampling distances (12–15 cm) and viewing angles.

In the following, Section 2 provides a brief overview of a number of existing VOT methods. Section 3 describes the proposed SMSOT-CNNs approach. Section 4 explains the experimental setup and discusses the results. Section 5 concludes the paper.

2. RELATED WORKS

Kalman and Particle filtering have been widely used in single object tracking. The approaches based upon Kalman filtering consider both the speed and position of motion resulting in accurate object tracking (Cuevas et al., 2005, Okuma et al., 2004).

(Bochinski et al., 2017) proposed a simple tracking approach based on computing Intersection-over-Union (IoU) in matching an object in consecutive frames by overlapping the frames. As is uses the position information, it works on very high speed, however, sacrificing the accuracy particularly in complex objects and scenes. Another similar method which uses position information is Simple Online and Real-time Tracking (SORT) method as an online tracker (Bewley et al., 2016) which can be combined with CNN-based object detectors (Azimi et al., 2018,

[1] https://www.ipf.kit.edu/code.php

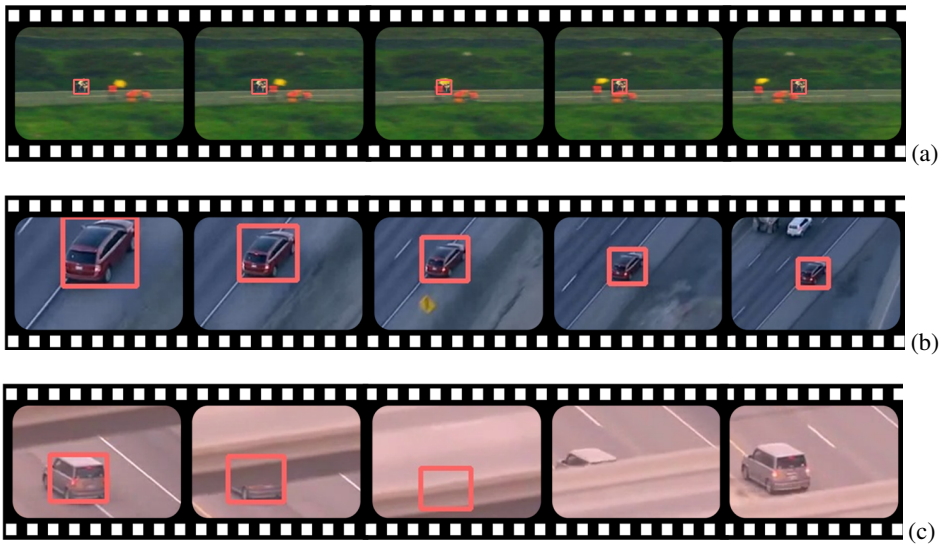


Figure 3. Indication of the tracking results of GOTURN on example aerial image sequences acquired by web search. (a) Successful tracking with small object and the presence of occlusion. (b) Tracking the object in the presence of scale change between the frames. (c) Example of a failure case when the object is occluded for a longer time.

?). It employs Kalman filter for predicting the object positions in each frame based on their positions in the previous frame. Hungarian method is used for matching the detected objects in the frame with the predicted positions. Finally, IoU of the detected and predicted bounding boxes is used as affinity measure for matching the detection and the track. SORT methods are more accurate and precise than the simple IoU trackers; nevertheless, they introduce more false positives. Deep SORT (Wojke et al., 2017) adds re-identification to affinity between the detections and tracks which increases the tracking accuracy. A combination of appearance, motion, and interaction affinity components through Recurrent Neural Networks (RNNs) shows promising tracking results (Sadeghian et al., 2017). In this approach, deep learning networks are used in online MOT scenario both for object detection and affinity modeling including appearance modeling by re-identification approaches. Approaches relaying on Correlation Filters (CF) have recently shown outperforming the key point-matching methods for multiple object tracking (Kart et al., 2019). A combination of CNN-based features with CF trackers has shown superior tracking accuracy in several tracking benchmarks (Milan et al., 2016) although using the deep CNN features increases the computational costs.

MDNet (Multi-Domain Network) takes advantage of the domain-specific and shared layers in order to track different objects from various domains. While the shared layers are trained offline, the domain-specific layers are trained online on the target video frames. The network is trained iteratively, where in each iteration, one domain-specific branch is trained on a video. In spite of its high tracking accuracy, MDNet is rather slow in processing the videos with high-frame rates and therefore, it is used only when the high accuracy is required.

GOTURN (Tracking Using Regression Networks) was proposed to alleviate the speed issue while preserving the tracking accuracy. The CNN layers of GOTURN are trained on collections of images and videos with bounding box annotations. In the inference time, GOTURN is applied with the frozen weights and without fine-tuning which allows it to reach the high speed of 100 fps. Figure 2 (a) shows the overview of GOTURN. Using the light-weight CNNs together with the offline training and online tracking, GOTURN can accurately track single objects in a high frame rate. Thus in this work, we use it in designing our SMSOT-CNNs approach.

3. METHODOLOGY

In the proposed SMSOT-CNNs approach, we assign each object with its given initial location to a MSOT-CNN which will be responsible for tracking the object through the image sequence. Then we integrate the results of all MSOT-CNNs for each image frame to derive the final multi-object tracking result of the frame. As a base CNN for designing the SMSOT-CNNs, we consider the GOTURN network (Held et al., 2016). This method has shown to be significantly faster than the other CNN-based methods. It considers less number of candidate object locations and therefore, provides course object localization which is refined through a bilinear interpolation of the extracted CNN feature maps from the course localization phase. Utilizing interpolations instead of extracting new feature maps (for finer localization) speeds up the tracking process significantly.

In order to validate the tracking performance of GOTURN on aerial image sequences, we selected three different videos with the frame rate of 25–35 fps from the web. The videos represent different challenges such as changes in viewpoints, scales, and illuminations together with the presence of occlusion and background clutter. Figure 1 demonstrates examples of the tracking results in which for a better illustration of the results, we have sampled and presented representative frames. As the results show, GOTURN is able to detect small objects even when they are partially occluded (Figure 1 a). Furthermore, it is able to keep track of the object even if the scale changes throughout the image sequences (Figure 1 b). However, in some cases such as when the object is occluded for a longer time, GOTURN may fail tracking the object (Figure 1 c).

GOTURN takes two consecutive video frames as input to the neural network modules and predicts the tracked object location in the current frame. We train it entirely offline with video frames and images so that it learns a generic relationship between the appearance and motion of objects and use it for tracking new objects in the inference time. In order to show

Input (previous frame) 227x227x3	Input (next frame) 227x227x3
Conv11-96-4	Conv11-96-4
Max-pooling	Max-pooling
Conv5-256-1	Conv5-256-1
Max-pooling	Max-pooling
Conv3-384-1	Conv3-384-1
Conv3-384-1	Conv3-384-1
Conv3-256-1	Conv3-256-1
Max-pooling	Max-pooling
Concat	
FC-4096	
FC-4096	
FC-4096	
FC-4	

(a) GOTURN

Input (previous frame) 227x227x3	Input (next frame) 227x227x3
Conv11-96-4	Conv11-96-4
Max-pooling	Max-pooling
Conv5-256-1	Conv5-256-1
Max-pooling	Max-pooling
Conv3-384-1	Conv3-384-1
Conv3-384-1	Conv3-384-1
Conv3-256-1	Conv3-256-1
Max-pooling	Max-pooling
Concat	
Conv3-512-1	
FC-4096	
FC-4096	
FC-4096	
FC-4	

(b) GOTURN+CONV

Input (previous frame) 227x227x3	Input (next frame) 227x227x3
Conv11-96-4	Conv11-96-4
Max-pooling	Max-pooling
Conv5-256-1	Conv5-256-1
Max-pooling	Max-pooling
Conv3-384-1	Conv3-384-1
Conv3-384-1	Conv3-384-1
Conv3-256-1	Conv3-256-1
Max-pooling	Max-pooling
Concat	
Conv3-512-1	
Conv3-512-1	
Conv3-512-1	
FC-4096	
FC-4096	
FC-4096	
FC-4	

(c) GOTURN+3CONV

Figure 4. Network structures of GOTURN and our two modifications to it.

the object of interest to the network, we scale a crop of the previous frame centered on the object and give it to the network as input. Afterwards, we pad the crop to provide the network with some contextual information of the object's surroundings. Assuming that the object moves smoothly through space, we select and crop a search region in the current frame based on the object's location in the previous frame. Then we scale the search region and feed it to the network as the second input. Afterwards, we regress via the network the object coordinates i.e. bottom right and top left corners of the corresponding bounding box in the search region. Figure 2 (a) represents the overview of the GOTURN method.

The GOTURN architecture composed of two identical sequences of convolutional layers that extract the features of the input image crops at different levels. The outputs of these convolutional layers are concatenated and fed into a sequence of fully connected (FC) layers. These layers compare the features of the object of interest in the previous frame to the features of the search region in the current frame in order to find the new position of the object. Figure 4 (a) details the GOTURN's architecture, where X, Y, and Z in each ConvX-Y-Z denote the convolutional layer's kernel size, number of output filters, and stride value, respectively. In GOTURN, the FC layers should learn complex features to be robust against variety of changes occurring to the object (e.g., translation, rotation, illumination, occlusion, deformation) between the two frames. The last FC layer comprise 4 nodes representing the coordinates of the output bounding box. After each convolutional and FC layer, ReLU non-linearity is used. In addition, a dropout non-linearity is applied to each FC layer.

In order to increase the tracking accuracy of GOTURN, we propose a number of modifications to the network structure such as applying batch normalization after the first convolutional layers and adding extra convolutional layers before the FC layers. Figure 2 (b) and Figure 4 (b,c) demonstrate the GOTURN network structure modified by adding extra convolutional layers. The ablation study of the networks on the KIT AIS dataset is provided in Section 4.

4. RESULTS AND DISCUSSION

For our experiments, we use Titan XP GPUs. The training configuration is similar to the original GOTURN (Held et al., 2016), with a scheduled learning rate of 10e-6 and the SGD optimizer. In order to quantitatively evaluate the trained models, we consider the widely-used metrics in the MOT domain (Milan et al., 2016) listed in Table 1. Among all these metrics, the MOT performance can be generally evaluated by MOTA and MOTP. MOTA considers all tracking errors including false positives, misses, and mismatches throughout all frames.

Metrics	Descriptions
IDF1	ID F1-Score
IDP	ID Global Min-Cost Precision
IDR	ID Global Min-Cost Recall
TP	True Positive – Number of Detected Objects
FP	False Positive – Number of False Detections
FN	False Negative – Number of Lost Objects
Rcll	Recall – TP over Number of Objects
Prcn	Precision – TP over Sum of TP and FP
FAR	False Acceptance Rate
MT	Ratio of Mostly Tracked Trajectories
PT	Ratio of Partially Tracked Trajectories
ML	Ratio of Mostly Lost Trajectories
IDS	Number of Identity Switches
FM	Fragmentation
MOTA	Multiple Object Tracker Accuracy
MOTP	Multiple Object Tracker Precision
MOTAL	Multiple Object Tracker Accuracy Log
Hz	Tracker Speed in Frame per Second

Table 1. Descriptions of the metrics used for quantitative evaluations.

$$MOTA = 1 - \frac{\Sigma_t(FN_t + FP_t + IDS_t)}{\Sigma_t GT_t}, \quad (1)$$

where t is the frame index and GT denotes the number of ground truth objects. MOTP evaluates the trackers' precision in estimating object positions. It is computed as the total position error for the matched objects throughout all frames averaged by the total number of matched objects.

$$MOTP = \frac{\Sigma_{t,i} d_{t,i}}{\Sigma_t c_t}, \quad (2)$$

In these evaluations, we consider three categories of tracked objects: Mostly Tracked (MT) – object is tracked successfully for $>$80% of its lifetime, Mostly Lost (ML) – object is tracked successfully for $<$20% of its lifetime, and Partially Tracked (PT) which are the rest of the objects.

The KIT AIS dataset composed of 9 vehicle tracking sequences with 229 frames and 13 person tracking sequences with 190

frames. The images are acquired by DLR-3K camera system composed of a nadir-looking and two side-looking DSLR cameras, mounted on an airborne platform. The image sequences were recorded with the frame rate of 2 Hz, and with different ground sampling distances (12–15 cm) and viewing angles. The length of the sequences in the vehicle dataset ranges from 14 to 47 frames, and in the person dataset ranges from 4 to 24 frames. The person dataset is split into a training and an evaluation set with 7 and 9 image sequences, respectively. However, the vehicle dataset is provided as a whole. Therefore, for our experiments, we split it into a training and evaluation set with 5 and 4 image sequences, respectively.

Table 2 shows the ablation study of GOTURN and its modified variants on the vehicle sequences of KIT AIS dataset, using different configurations. According to the table, original GOTURN (GOT) achieves MOTA of 23.0 and MOTP of 73.3 with the batch size of 50. Increasing the batch size to 150 improves MOTA and MOTP to 24.7 and 73.6, respectively. This indicates the key role of the batch size in training a CNN network in the tasks with large number of objects. In the training of GOT, only the FC layers were trained whereas the convolutional layers were frozen with the weights trained on the ImageNet dataset, similar to the original work (Held et al., 2016). Nevertheless, in our experiments, the images and therein objects look significantly different from those of in the ImageNet dataset. Therefore, by training GOT with not frozen (NF) convolutional weights, we can observe dramatic increase of MOTA and MOTP to 32.0 and 77.1, respectively.

Recently, batch normalization (BN) layer has shown superior performance in comparison with dropout layers as a regularization technique avoiding overfitting issues (Ioffe, Szegedy, 2015). In order to evaluate the effects of BN on the tracking performance of GOTURN, we add a single BN layer after the first convolutional layers in both branches. However, according to the results, this ruins the tracking accuracy (MOTA drops to 19.2). We suppose that BN layers receive very diverse regions of interest and thus, normalization dismisses the unique features for each object of interest and merges their features. More comprehensive investigations are needed to find out the reason behind the network behaviour in the presence of BN.

In order to make the network deeper and extract richer fused features, we insert extra convolutional layers after concatenating the features from the two branches. We consider adding one and three convolutional layers separately to investigate the network's depth impact on the tracking performances. According to the results, even adding a single convolutional layer significantly boosts the tracking accuracy and increases MOTA from 32.0 to 39.1. Increasing the number of extra convolution layers to three, further improves MOTA to 41.1. This verifies the significance of richer high-level fused features in the tracking accuracy. According to Table 2, although increasing the number of additional convolutional layers from one to three generally increases the tracking accuracy, it slightly reduces the percentage of the mostly tracked (MT) objects. This could be due to the ID switch or fragmentation. Hints that the stack of convolutional layers after the concatenation step can increases the risk of losing track of objects.

Table 3 represents the ablation study results of person tracking on the KIT AIS dataset. The results indicate that the network behaves similarly for the vehicle and person tracking scenarios. In training the networks with unfrozen convolutional weights, we used the DLR's Aerial Crowd Dataset (Bahmanyar et al., 2019) as an additional aerial image dataset to increase the diversity of the learned features. As it can be seen in Table 3, the values of MOTA and MOTAL are negative. Due to the large number of people in the sequences and the complexities of the scenes, there is a higher probability of missing people which results in many lost bounding boxes all over the scenes, i.e., large FP and FN. Comparing the results of Tables 2 and 3 demonstrate that the person tracking in aerial imagery is a more complex than vehicle tracking for GOTURN, as all the metrics shows significant lower performance of the tracker on the person dataset.

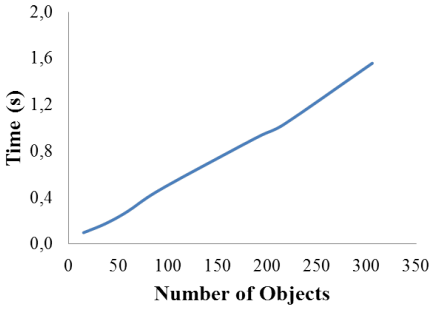


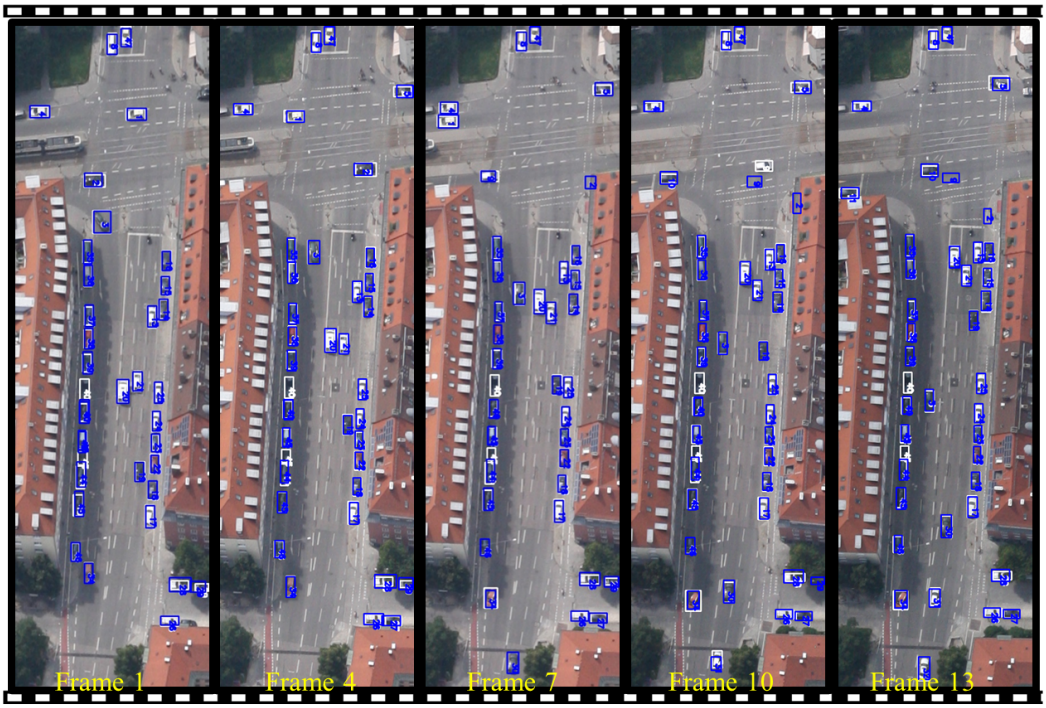
Figure 5. Linear increasing of the inference time for each frame by the increase in the number of objects in the frames.

Tables 4 and 5 show the tracking results of the best performing model (GOT-NF-3Conv) on the test image sequences of the vehicle and person datasets. The results show that most of the errors in both datasets are due to fragmentation and FPs. This could be caused by occlusions and complex backgrounds in the scenes. We suppose that training the models on a larger dataset with a more scene diversity could help overcoming these issues to a large degree. Figures 6 and 7 illustrate examples of the vehicle and person tracking results, where the white and blue bounding boxes depict the ground truth and predictions, respectively.

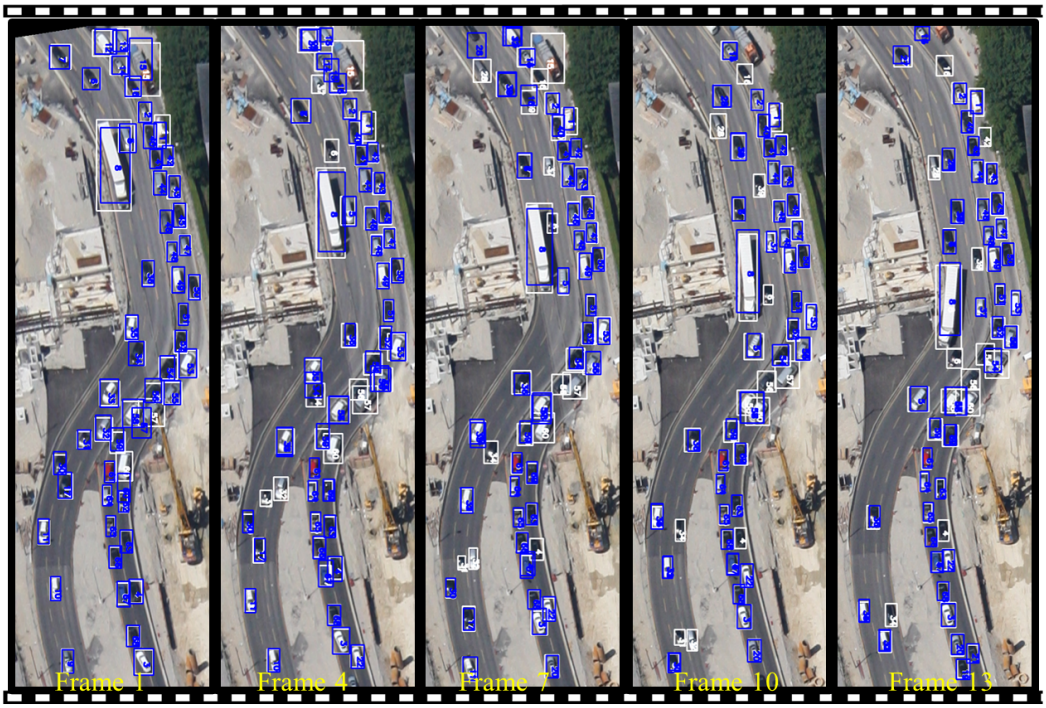
Figure 5 represents the required processing time of the SMSOT-CNNs approach versus the number of objects presented in the scenes. According to the figure, the complexity is increased linearly by the increase in the number of tracked objects.

5. CONCLUSION

In this work, we proposed a new approach based on a stack of micro CNNs to track multiple vehicles and people in aerial image sequences. The proposed method is able to make use of the promising performance of CNNs while keep the computation cost reasonable in the presence of large number of objects in aerial images. As a base single-object tracker based on CNN, we selected the GOTURN network and validated its performance on the vehicle and person sets of KIT AIS tracking dataset. In addition, we modified the GOTURN network by adding a number of convolutional layers in order to enrich its higher level features. Results demonstrated that this modification leads to 28.4% and 10.2% increase in the vehicle and person tracking performance, respectively. Due to the existing occlusions and complex backgrounds, a larger and more diverse dataset could help improving the tracking results in future studies. Furthermore, other state-of-the-art light-weight tracking networks could be considered as the base tracking method.



(a)



(b)

Figure 6. Sample illustrations of the tracking results for the vehicle dataset based on GOT-NF-3Conv/150 model.

Model/batch size	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	FAR↑	MT (%)↑	PT (%)↓	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA↑	MOTP↑	MOTAL↑
GOT/50	60.8	59.1	62.5	64.9	61.3	18.9	56.5	24.3	19.2	2037	1748	47	91	23.0	73.3	23.9
GOT/150	61.2	59.7	62.9	65.4	62.0	18.4	59.1	20.0	20.9	1991	1720	34	99	24.7	73.6	25.4
GOT-NF/150	65.6	63.7	67.7	69.3	65.3	16.98	64.3	19.5	16.2	1834	1525	25	63	32.0	77.1	32.4
GOT-NF-BN/150	59.2	57.5	61.0	63.1	59.5	19.79	58.3	17.8	23.9	2137	1835	47	70	19.2	76.2	20.1
GOT-NF-Conv/150	69.4	67.5	71.3	72.5	68.7	15.22	**70.0**	**15.2**	14.8	1644	1367	16	58	39.1	77.1	39.4
GOT-NF-3Conv/150	**69.8**	**68.2**	**71.5**	**73.1**	**69.7**	**14.6**	69.1	17.4	**13.5**	**1577**	**1338**	**14**	57	**41.1**	**77.3**	**41.4**

Table 2. Quantitative results of multi object tracking ablation study for the Vehicle dataset with the total number of 230 vehicles. The sequences are of 4.5 Hz. ↑ and ↓ stand for "higher and lower is better" accordingly.

Model/batch size	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	FAR↑	MT (%)↑	PT (%)↓	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA↑	MOTP↑	MOTAL↑
GOT/32	25.2	24.5	25.8	30.7	29.1	130.1	13.8	54.6	31.6	8976	8340	297	703	-46.5	67.9	-44.0
GOT/150	27.5	26.8	28.2	33.2	31.6	125.3	15.2	55.4	29.4	8649	8031	275	721	-41.0	67.5	-38.7
GOT-NF/150	32.3	31.5	33.1	36.8	35.0	119.1	23.4	52.4	24.2	8213	7605	198	619	-33.2	70.0	-31.6
GOT-NF-BN/150	31.1	30.4	31.9	35.4	33.7	121.6	20.0	57.0	23.0	8393	7767	167	621	-35.8	69.6	-34.4
GOT-NF-Conv/150	33.6	32.8	34.5	37.9	35.9	117.7	24.9	**51.8**	23.3	8122	7474	171	620	-31.1	70.5	-29.7
GOT-NF-3Conv/150	**34.0**	**33.2**	**34.9**	**38.2**	**36.4**	**116.4**	**25.0**	52.5	**22.5**	**8028**	**7427**	**157**	**614**	**-29.8**	**71.0**	**-28.5**

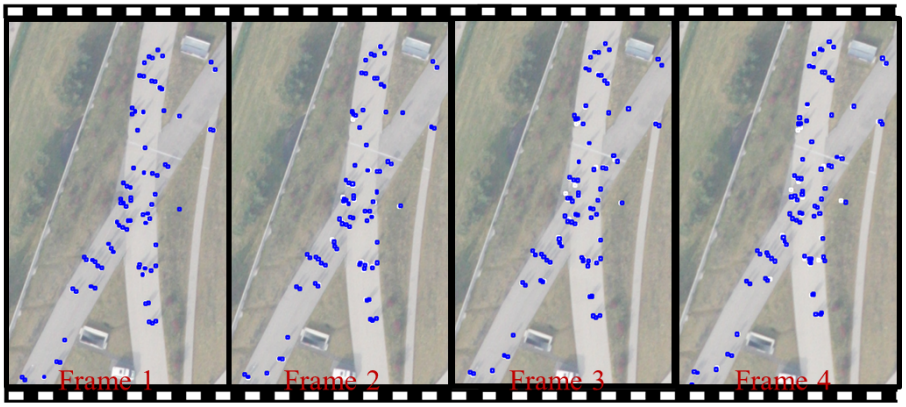
Table 3. Quantitative results of our multi object tracking ablation study for the Person dataset with the total number of 1043 people. The sequences are of 1.1 Hz. ↑ and ↓ stand for "higher and lower is better" accordingly.

Image Sequence	# Images	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	FAR↑	GT	MT (%)↑	PT (%)↓	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA↑	MOTP↑	MOTAL↑
MunichCrossroad-02	46	57.4	56.2	58.6	60.4	58.0	20.96	66	54.5	24.2	21.3	943	854	7	32	16.3	72.5	16.6
MunichStreet-02	21	89.0	85.8	92.4	92.5	85.9	5.6	47	91.5	2.1	6.4	113	56	0	4	77.3	81.8	77.3
MunichStreet-04	23	78.0	76.9	79.16	80.6	78.3	11.7	68	75.0	14.7	10.3	339	294	3	9	58.1	80.9	58.3
StuttgartCrossroad-01	15	69.7	66.9	72.7	75.8	69.8	13.0	49	59.2	26.5	14.3	182	134	4	12	42.2	74.4	42.8
overall	105	69.8	68.2	71.5	73.1	69.7	14.6	230	69.1	17.4	13.5	1577	1338	14	57	41.1	77.3	41.4

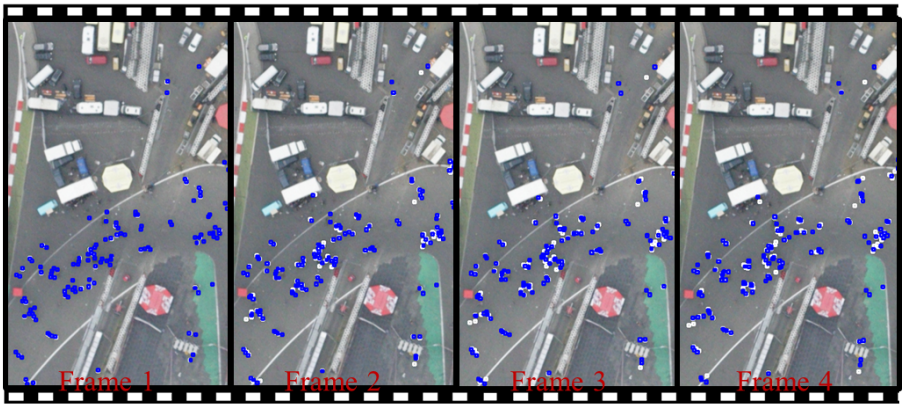
Table 4. Quantitative results of the GOT-NF-3Conv/150 model on different image sequences of the Vehicle dataset. ↑ and ↓ stand for "higher and lower is better" accordingly.

Image Sequence	# Images	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	FAR↑	GT	MT (%)↑	PT (%)↓	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA↑	MOTP↑	MOTAL↑
AA-Crossing-02	13	49.9	49.7	50.1	52.1	51.6	42.6	94	24.5	52.1	23.4	554	544	11	71	2.3	68.8	3.2
AA-Walking-02	17	30.7	30.2	31.3	33.8	32.7	109.6	188	15.5	38.9	45.6	1864	1767	34	140	-37.2	68.0	-36.0
Munich-02	31	23.6	22.7	24.5	28.8	26.7	156.3	230	8.6	38.3	53.1	4846	4363	105	316	-52.1	68.4	-50.4
RaR-Snack-Zone-02	4	61.6	61.4	61.8	64.4	63.9	78.5	220	37.3	62.3	0.4	314	308	2	39	27.9	77.9	28.0
RaR-Snack-Zone-04	4	61.2	61.1	61.3	63.8	63.6	112.5	311	34.4	64.6	1.0	450	445	5	48	26.8	76.7	27.2
Overall	69	34.0	33.2	34.9	38.2	36.4	116.4	1043	25.0	52.5	22.5	8028	7427	157	614	-29.8	71.0	-28.5

Table 5. Quantitative results of the GOT-NF-3Conv/150 model on different image sequences of the Person dataset. ↑ and ↓ stand for "higher and lower is better" accordingly.



(a)



(b)

Figure 7. Sample illustrations of the tracking results for the vehicle dataset based on GOT-NF-3Conv/150 model.

REFERENCES

Azimi, S. M., Vig, E., Bahmanyar, R., Körner, M., Reinartz, P., 2018. Towards multi-class object detection in unconstrained remote sensing imagery. *ACCV*.

Bahmanyar, R., Vig, E., Reinartz, P., 2019. MRCNet: Crowd counting and density map estimation in aerial and ground imagery. *In BMVC's Workshop on Object Detection and Recognition for Security Screening (BMVC-ODRSS)*.

Benfold, B., Reid, I., 2011. Stable multi-target tracking in real-time surveillance video. *CVPR*, IEEE.

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 3464–3468.

Bochinski, E., Eiselein, V., Sikora, T., 2017. High-speed tracking-by-detection without using image information. *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 1–6.

Cuevas, E. V., Zaldivar, D., Rojas, R., 2005. Kalman filter for vision tracking.

El-Shafie, A., Zaki, M., Habib, S., 2019. Fast CNN-based object tracking using localization layers and deep features interpolation. *In arXiv prepreint arXiv:1901.02620v1*.

Girshick, R., 2015. Fast R-CNN. *CVPR*.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*.

Held, D., Thrun, S., Savarese, S., 2016. Learning to track at 100 fps with deep regression networks. *ECCV*.

Henriques, J. F., Caseiro, R., Martins, P., Batista, J., 2014. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3), 583–596.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kart, U., Lukezic, A., Kristan, M., Kamarainen, J.-K., Matas, J., 2019. Object tracking by reconstruction with view-specific discriminative correlation filters. *CVPR*.

Kristan, M., Matas, J., Leonardis, A., Vojíř, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L., 2016. A novel performance evaluation methodology for single-target trackers. *IEEE transactions on pattern analysis and machine intelligence*, 38(11), 2137–2155.

Lin, T., Dollár, P., Girshick, R. B., He, K., Hariharan, B., Belongie, S. J., 2017. Feature Pyramid Networks for Object Detection. *CVPR*.

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Zhao, X., Kim, T.-K., 2014. Multiple object tracking: A literature review. *arXiv preprint arXiv:1409.7618*.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K., 2016. MOT16: A Benchmark for Multi-Object Tracking. *arXiv:1603.00831 [cs]*.

Nam, H., Han, B., 2016. Learning multi-domain convolutional neural networks for visual tracking. *CVPR*, 4293–4302.

Okuma, K., Taleghani, A., De Freitas, N., Little, J. J., Lowe, D. G., 2004. A boosted particle filter: Multitarget detection and tracking. *ECCV*, Springer.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS*.

Sadeghian, A., Alahi, A., Savarese, S., 2017. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. *ICCV*, 300–311.

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, 3645–3649.