UA-DETRAC 2018: Report of AVSS2018 & IWT4S Challenge on Advanced Traffic Monitoring

Siwei Lyu[1,14], Ming-Ching Chang[1], Dawei Du[1], Wenbo Li[1], Yi Wei[1], Marco Del Coco[2], Pierluigi Carcagnì[2], Arne Schumann[3], Bharti Munjal[8], Dinh-Quoc-Trung Dang[9], Doo-Hyun Choi[7], Erik Bochinski[10], Fabio Galasso[8], Filiz Bunyak[12], Guna Seetharaman[13], Jang-Woon Baek[6], Jong Taek Lee[6], Kannappan Palaniappan[12], Kil-Taek Lim[6], Kiyoung Moon[6], Kwang-Ju Kim[6], Lars Sommer[4,3], Meltem Brandlmaier[8], Min-Sung Kang[5], Moongu Jeon[11], Noor M. Al-Shakarji[12], Oliver Acatay[3], Pyong-Kun Kim[6], Sikandar Amin[8], Thomas Sikora[10], Tien Dinh[9], Volker Eiselein[10], Vu-Gia-Hy Che[9], Young-Chul Lim[5], Young-min Song[11], and Yun-Su Chung[6]

[1]University at Albany, State University of New York, USA, [2]National Research Council, Italy, [3]Fraunhofer IOSB, Germany, [4]Karlsruhe Institute of Technology, Germany, [5]DGIST, South Korea, [6]Electronics and Telecommunications Research Institute, South Korea, [7]Kyungpook National University, South Korea, [8]OSRAM GmbH, Germany, [9]Ho Chi Minh University of Science, Vietnam, [10]Technische Universität Berlin, Germany, [11]Gwangju Institute of Science and Technology, South Korea, [12]University of Missouri Columbia, USA, [13]U.S Naval Research Laboratory, USA, [14]Tianjin Normal University, China

Abstract

A desirable smart traffic-monitoring and street-safety system can elicit and support the intervention of law enforcement agencies or medical staff. Recently, there has been a dramatically higher demand for such smart systems. To this end, the International Workshop on Traffic and Street Surveillance for Safety and Security (IWT4S) was organized in conjunction with the 15th IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS 2018). Our goal is to advance the state-of-the-art detection and tracking algorithms and provide a comprehensive performance evaluation for them. We evaluate 5 *submitted detection and* 7 *submitted tracking methods on the large-scale UA-DETRAC benchmark, and the results are shared publicly on the website* http://detrac-db.rit.albany.edu. *We expect this challenge to advance the research and development of new detection and tracking methods for transportation applications.*

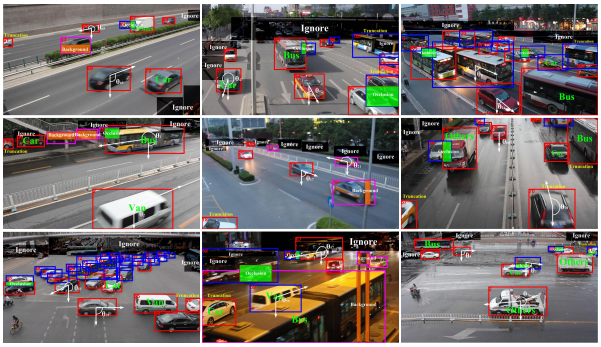

Figure 1. Examples of annotated frames in the UA-DETRAC datasets. Colors of the boundary of the bounding boxes reflect the occlusion property, as fully visible (red), partially occluded by other vehicles (blue), or partially occluded by objects from the site (pink). Black opaque regions are ignored, green opaque boxes are regions occluded by other vehicles, and orange opaque boxes are regions occluded by objects from the site. White arrow over the vehicle shows its orientation angle.

1. Introduction

With the advent of ubiquitous smart camera systems, traffic surveillance becomes the major method that is low-cost and effective to alleviate the inefficient and ineffective transportation systems. The basic expectation for a desirable traffic surveillance system is to provide the accurate and consistent trajectory of objects of interest in the scene. Therefore, effective object detection and multi-target tracking methods are the basic building blocks for a reliable traffic surveillance system.

The past few years have witnessed a great deal of progress on object detection performance, thanks to a number of state-of-the-art visual detection algorithms [18, 13, 5, 16, 2, 21, 24]. Leveraging the advance in detections, track-

ers based on the tracking-by-detection paradigm have also achieved significantly better performance [3, 23, 11].

Traffic surveillance based on visual detection and tracking is still considered unsolved due to several factors, *e.g.*, scale changes, occlusions, and background clutters. Apart from these difficulties, more insights can be found in the previous IWT4S challenge held at Leece, Italy [14]: The best available methods still yield mediocre performance on two widely used metrics, *i.e.*, ID switches (PR-IDS) and fragmentation (PR-FM), not to mention the additional complications and requirements that will arise when applying to real-world traffic surveillance use cases.

The use of deep learning and deep neural networks in detection and tracking continues to dominate the front, which accelerates the development with fast pace, thanks to many breakthroughs including the low-cost high-performance GPUs. Evaluation and benchmark of the state-of-the-art methods on a large-scale dataset will enable a fair 'comparison of apples and oranges' and provide valuable insight for future developments. Is it true that the object detection performance will determine the upper bound of the tracker performance? Will hybrid methods outperform detect-by-tracking? To answer these questions, we held the *AVSS2018 Challenge on Advanced Traffic Monitoring* in conjunction with the *International Workshop on Traffic and Street Surveillance for Safety and Security* (IWT4S) using the large-scale UA-DETRAC benchmark [22] (Figure 1) as the evaluation platform. This paper summarizes the challenge results, with observations and discussions that can shed lights on the next generation of traffic video analysis advancements.

2. AVSS2018 Challenge

This challenge is a continuation from last year, the *AVSS2017 Challenge on Advanced Traffic Monitoring* [14], with an aim to to find the most powerful object detection and tracking methods for traffic and street video analysis. Again the UA-DETRAC benchmark evaluation [22] is used, which consists of 100 video sequences, which is divided into a training set (60 sequences) and a test set (40 sequences). There are more than $140,000$ frames in the UA-DETRAC benchmark that are manually annotated corresponding to $8,250$ different vehicles. Participant teams are allowed to use additional training data to optimize their performance without constraint.

Similar to [14], teams can participate the challenges in two difficulty levels, *i.e.*, the *beginner* and *experienced* levels. The *beginner* level is based on the "easy" subset (10 sequences) in the UA-DETRAC benchmark [22], while the *experienced* level contains the whole test set (40 sequences) for detection or the test-medium and test-hard set (30 sequences) for tracking. Besides, we further analyze the performance of submitted algorithms under the 4 weather conditions provided in the UA-DETRAC dataset, *i.e.*, *cloudy*, *night*, *sunny*, and *rainy*. In terms of evaluation, we use the Average Precision (AP) score [19], the precision-recall curve for tracking and DETRAC metrics [22] for tracking. Refer to the challenge website[1] for more information.

Table 1. Average running speed of the detection algorithms (in FPS) when running on the UA-DETRAC-test set. "×" indicates that GPU is not used. "n/a" indicates that the result is not provided by the participant team.

Detectors	GPU	RAM	Implementation	Speed
RD^2 (A.1)	Tesla P40	n/a	Python	n/a
ExtendNet (A.2)	TitanX	32GB	C/C++	45.45
MSVD_SPP (A.3)	TitanX	64GB	Python	n/a
IMIVD-TF (A.4)	n/a	n/a	Tensorflow	1
MYOLO (A.5)	n/a	n/a	C/C++	7
DPM [7]	×	8GB	Matlab,C++	0.17
ACF [6]	×	64GB	Matlab	0.67
R-CNN [8]	Tesla K40	64GB	Matlab,C++	0.10
CompACT [4]	Tesla K40	64GB	Matlab,C++	0.22
GP-FRCNN [14]	Tesla K40	256GB	Python, C++	4.00

3. T4S 2018 Performance Evaluation

In this section, we compare all submitted methods this year against the baseline and the best performers in the T4S2017 Challenge. We start with overview of the submission methods and then present the evaluation results.

3.1. Object Detection

We received 5 object detection submissions in the T4S 2018 challenge (1 submission in the beginner level, and 5 submissions in the experienced level). The submitted methods are evaluated against the winner GP-FRCNN [14] in the T4S 2017 challenge and 4 baseline detection methods (*i.e.*, DPM [7], ACF [6], R-CNN [8], and CompACT [4]). Thus, totally 10 methods are included in the T4S 2018 detection challenge. Refer to Appendix A for the technical details of the submission methods and Table 1 for speed comparison.

All submissions are improvements from the state-of-the-art deep learning object detection methods, including Faster R-CNN [17] (*e.g.*, IMIVD-TF (A.4)), YOLOv3 [16] (*e.g.*, MSVD_SPP (A.3) and MYOLO (A.5)), and RefineDet [24] (*e.g.*, RD^2 (A.1)). Moreover, the results show good performance of methods such as the SENet [10] and SPP Network.

Table 2 summarizes the detection benchmark of this T4S2018 challenge. All submitted methods outperform the GP-FRCNN [14], which is the winner of the T4S 2017 challenge, by at least 7% in the AP score. We found that such an improvement mainly comes from better performance in difficult weather conditions, *i.e.*, *rainy* and *night*.

In the beginner level, the best submitted method is RD^2 (A.1), which achieves 96.03% AP score with a 5% improve-

[1] https://iwt4s2018.wordpress.com/

Table 2. AP scores of submitted object detection algorithms on the UA-DETRAC-test set in the beginner / experienced levels in various environmental conditions. "−" indicates the data is not available. Bold faces are the top performers.

Detectors	Overall	Easy	Medium	Hard	Cloudy	Night	Rainy	Sunny
RD^2 (A.1)	**96.03**/85.35	**96.03**/95.80	−/89.84	−/**76.64**	**98.64**/**89.67**	**96.24**/86.59	**88.30**/78.17	**95.03**/90.49
ExtendNet (A.2)	−/83.59	−/95.46	−/88.75	−/73.36	−/86.89	−/85.05	−/76.75	−/**90.77**
MSVD_SPP (A.3)	−/84.96	−/95.59	−/89.95	−/75.34	−/88.12	−/88.81	−/77.46	−/89.46
IMIVD-TF (A.4)	−/**85.67**	−/**96.32**	−/**91.17**	−/75.45	−/87.02	−/**88.93**	−/**80.60**	−/89.69
MYOLO (A.5)	−/83.50	−/95.15	−/88.18	−/73.99	−/88.58	−/83.38	−/77.06	−/88.37
GP-FRCNN [14]	91.90/76.57	91.90/91.79	−/80.85	−/66.05	92.77/81.23	92.91/77.20	82.77/68.59	93.96/85.16
DPM [7]	34.63/25.70	34.63/34.42	−/30.29	−/17.62	32.54/24.78	36.71/30.91	40.26/25.55	50.53/31.77
ACF [6]	54.80/46.35	54.80/54.27	−/51.52	−/38.07	71.95/58.30	45.20/35.29	43.76/37.09	73.07/66.58
R-CNN [8]	59.71/48.95	59.71/59.31	−/54.06	−/39.47	74.14/59.73	94.46/39.32	90.81/39.06	76.64/67.52
CompACT [4]	65.50/53.23	65.50/64.84	−/58.70	−/43.16	77.27/63.23	61.98/46.37	57.68/44.21	77.35/71.16

Table 3. PR-DETRAC metrics of tracking algorithms on the UA-DETRAC-test set in the beginner/ experienced levels"−" indicates the data is not available. Bold faces are the top performers.

Trackers	Detection	PR-MOTA↑	PR-MOTP↑	PR-MT↑	PR-ML↓	PR-IDS↓	PR-FM↓	PR-FP↓	PR-FN↓
GOG [15]	CompACT [4]	23.9/11.7	**47.4**/34.4	20.5%/10.8%	21.0%/21.1%	829.9/2571.2	776.2/2463.8	6276.5/25352.8	36738.3/145257.5
IOUT [3]	EB [21]	34.0/16.4	37.8/26.7	27.9%/14.8%	20.4%/18.2%	573.6/1743.2	603.7/1846.3	**1617.0**/12627.0	33760.8/136077.8
JTEGCTD [14]	CompACT [4]	28.4/14.2	47.1/34.4	23.1%/13.5%	18.3%/18.7%	69.4/415.3	260.6/1345.7	5034.0/26221.8	33093.8/**133867.4**
JDTIF (B.1)	GP-FRCNN [2]	−/28.0	−/41.8	−/34.2%	−/20.9%	−/697.5	−/3431.8	−/55801.3	−/150493.4
MFOMOT (B.2)	R-CNN [8]	34.6/14.8	46.6/35.6	30.2%/11.9%	12.0%/20.8%	210.6/870.0	477.0/2035.2	3828.3/21277.4	27232.5/151788.2
KF-IOU (B.4)	RD^2 (A.1)	40.1/**31.0**	49.8/**49.9**	42.3%/**37.4%**	**5.8%**/**10.4%**	111.4/724.8	125.2/995.6	8674.4/52243.0	**13153.4**/**94728.1**
V-IOU (B.3)	FRCNN [17]	**37.9**/29.0	41.7/35.8	**38.1%**/30.1%	24.7%/22.2%	**18.7**/**142.0**	**39.8**/**244.0**	3855.1/**14177.0**	34738.5/143879.6
DMC (B.5)	CompACT [4]	−/14.6	−/34.1	−/11.6%	−/20.6%	−/908.3	−/1287.4	−/16056.7	−/141463.2
GMMA (B.6)	CompACT [4]	−/12.3	−/34.3	−/10.8%	−/21.0%	−/627.5	−/2423.7	−/25577.4	−/144148.9
SCTrack-3L (B.7)	CompACT [4]	25.9/12.1	47.2/35.0	15.0%/7.7%	20.6%/24.8%	91.8/378.3	323.7/947.5	2485.2/8241.0	38820.9/162937.6

ment compared to GP-FRCNN [14]. In the experienced level, IMIVD-TF (A.3) performs the best in most attributes among all the detection methods, which combines Faster R-CNN [17] with Neural Architecture Search (NAS) framework [25] as a base network. RD^2 (A.1) performs the best in the *rainy* and *night* conditions. This is attributed to the combination of RefineDet and SENet for more robustness. Notably, the ExtendNet (A.2) achieve fast detection speed of 22 ms, due to the combined forward of the network and NMS in the GPU processing.

3.2. Multi-Object Tracking

We received 7 object tracking submissions in the T4S 2018 challenge (4 submissions in the beginner level, and 7 submissions in the experienced level). The submitted methods are evaluated against with 3 best trackers (with the best detection input) from T4S 2017. Thus, a total of 10 trackers are evaluated in this challenge. Refer to Appendix B for the technical details of the submission methods and Table 4 for speed comparison.

The JDTIF (B.1) tracker performs joint detection and tracking by an end-to-end CNN model. The two trackers DMC (B.5) and SCTrack-3L (B.7) combine various features to increase discriminability. The four trackers, MFOMOT (B.2), V-IOU (B.3), KF-IOU (B.4), and GMMA (B.6), introduce the use of a single object tracking module to facilitate multiple object tracking.

Table 3 summarizes the performance of all submitted trackers. Compared to the best performer (with the CompACT detection) in the T4S 2017 challenge, the majority of the submitted methods achieves better accuracy in both the beginner and experienced levels. In the beginner level, V-IOU (B.3) achieves the best performance in terms of PR-MOTA, PR-MT, PR-IDS, and PR-FM, with the use of the state-of-the-art detector FRCNN [17]. In the experienced level, KF-IOU (B.4) performs the best in PR-MOTA, PR-MOTP, PR-MT and PR-ML metrics with the use of private detections from RD^2 (A.1). It can be concluded that better detection input generally achieves better tracking performance.

4. Discussions

Based on the above analysis, many submitted detectors and trackers perform significantly well against the baseline methods in T4S 2017 challenge. However, the best achieved AP score is 85.67% for detection and PR-MOTA score is 31.0 for tracking, which indicates that there is still room for improvement. We highlight some insights that might be useful for the improvement of the frontier of traffic video analysis, as well as the general visual object detection and tracking.

- **Model ensemble.** To deal with various scales of objects, a viable solution is to train different sub-models for different object scales. For example, MSVD_SPP (A.3) improve the YOLOv3 model [16] by adding two more object prediction layers to detect all sizes (large, medium, small) of objects.

- **Joint detection and tracking.** Currently, most object detection methods are designed for used in a still image. To deal with video sequences, detection and track-

ing modules can be learned jointly to exploit the spatio-temporal cues that can improve robustness. For example, JDTIF (B.1) proposes a new CNN model for the joint feature learning of a combined task of detection, tracking and identification.

- **Feature enhancement.** To improve the discriminability of similar objects, a recent approach is to combine multi-model features (*e.g.*, shape, color and deep features) that can provide a compact representation of target appearance. For example, DMC (B.5) employs multi-channel features (that includes three color channels and seven gradient channels) in order to calculate the similarity between the detections and tracklets.

- **Single object tracking for multiple object tracking.** Tracking-by-detection based multi-object methods heavily rely on the quality of input detections. When the detector fails, most tracking algorithms recover missing detections by a simple interpolation step that is performed within the trajectories. To reduce false negatives, one solution is to combine single object tracking with multiple object tracking that can construct a more discriminative appearance model to improve robustness (*e.g.*, KF-IOU (B.4) and GMMA (B.6)).

5. Conclusion

This paper summarizes the IWT4S 2018 challenge with evaluation results. Overall, 5 submitted detectors and 7 trackers are evaluated on the UA-DETRAC benchmark. The winners of the challenge are the following. The top detectors in the beginner and experienced levels are RD^2 (A.1) and IMIVD-TF (A.4), which achieves 96.03%, and 85.67% AP scores, respectively. The top trackers in the beginner and experienced levels are V-IOU (B.3) and KF-IOU (B.4), which achieves 37.9%, and 31.0% PR-MOTA scores, respectively. In general the 2018 T4S submissions are stronger than the 2017 contest, and there are still room for improvement. Insights and potential research directions are provided in the discussions. Looking forward, we plan to improve the UA-DETRAC benchmark with richer annotations that can be further used by additional real-world applications.

Acknowledgements. This work is partly supported by the National Science Foundation under Grant No. IIS-1537257, and the Nvidia Corporation.

A. Appendix A: Detection Submissions

A.1. Ensemble of two RefineDet models (RD^2)

Oliver Acatay, Lars Sommer, Arne Schumann
{oliver.acatay,lars.sommer,arne.schumann}@iosb.fraunhofer.de

RD^2 is a variant of the RefineDet [24] detector using the novel Squeeze-and-Excitation Network (SENet) [10] as base network. Two variants of the detector are trained: one with SEResNeXt-50 and one with ResNet-50 as base network, each with the same set of anchor sizes. The detection results of the two detectors are combined via averaging, where each detector is weighted equally.

Table 4. Average running speed (in FPS) of tracking algorithms on the UA-DETRAC-test set. "−" indicates that the data is not available, and "×" indicates that GPU is not used.

Trackers	Codes	CPU	RAM	Frequency	GPU	Speed
JDTIF (B.1)	-	E5-2690	256GB	2.60GHz	Quadro P6000	-
MFOMOT (B.2)	Python	i5-6200U	4GB	2.30GHz	×	-
V-IOU (B.3)	Python	i7-6700	32GB	-	×	-
KF-IOU (B.4)	Python	i7-2670QM	-	2.20GHz	×	-
DMC (B.5)	C++	i7-7740K	32GB	4.30GHz	×	-
GMMA (B.6)	C++	i7-7700K	32GB	4.20GHz	×	84.43
SCTrack-3L (B.7)	Matlab	i7-4720	16GB	2.60GHz	GTX960M	-
IOUT [3]	Python	i7-6700	32GB	3.40GHz	×	**6902.07**
JTEGCTD [14]	Matlab	i7-3720QM	8GB	2.70GHz	×	60.38
CEM	Matlab	i7-3520M	16GB	2.90GHz	×	4.62
GOG	Matlab	i7-3520M	16GB	2.90GHz	×	389.51
DCT	Matlab,C++	i7-3520M	16GB	2.90GHz	×	0.71
IHTLS	Matlab	i7-3520M	16GB	2.90GHz	×	19.79
H^2T	C++	i7-3520M	16GB	2.90GHz	×	3.02
CMOT	Matlab	i7-3520M	16GB	2.90GHz	×	3.79

A.2. Object Detection Network based on Single Extended Feature Map (ExtendNet)

Min-Sung Kang, Young-Chul Lim*
{mskang,linolyc}@dgist.ac.kr

ExtendNet is modified from the work in [12] published in conjunction with the 2018 IEEE Intelligent Vehicle Symposium. The paper describes a method for detecting three classes of street objects (cars, pedestrians, cyclists). This method is modified to detect the four types of vehicles in this challenge.

A.3. Multi-Stage Vehicle Detection with Spatial-Pyramid-Pooling (MSVD_SPP)

Kwang-Ju Kim, Pyong-Kun Kim, Yun-Su Chung, Doo-Hyun Choi*
{kwangju,iros,yoonsu}@etri.re.kr, dhc@ee.knu.ac.kr

MSVD_SPP is a multi-scale vehicle detection method with spatial pyramid pooling based on YOLOv3 [16]. Major improvements are summarized in the following. First, two more object prediction layers are introduced. Specifically, one additional prediction layer is added between the large-size and mid-size object perdition layers. The other prediction layer is added between the mid-size and small-size object prediction layers. Second, the Spatial Pyramid Pooling (SPP) networks are implemented before each prediction layer after the feature pyramid network.

A.4. Integrating Multiple Inference Instances of on Image for Vehicle Detection with Temporal Filtering (IMIVD-TF)

Jong Taek Lee, Jang-Woon Baek, Kiyoung Moon, Kil-Taek Lim
{jongtaeklee,jwbaek98,kymoon,ktl}@etri.re.kr

IMIVD-TF employs an unsupervised integration of multiple instances of an image by analyzing video sequences. Faster R-CNN [17] is applied with Neural Architecture Search (NAS)

framework [25] as a base network. Unsupervised integration of multiple instances of an image is employed by analyzing video sequences. Multiple cropped instances from images are generated, such that these instances become more similar to the images in the training set. After running the object detection algorithm on the additional instances, detection results from these runs and the source images are merged by non-maximum suppression (NMS).

A.5. Modified YOLO (MYOLO)

PyongKun Kim
iros@etri.re.kr
MYOLO is based on the original YOLO method with modifications in the anchor values *w.r.t.* the UA-DETRAC dataset.

B. Appendix B: Tracking Submissions

B.1. Joint Detection and Tracking in Videos with Identification Features (JDTIF)

Bharti Munjal, Sikandar Amin, Meltem Brandlmaier, Fabio Galasso
{m.bharti,s.amin,m.brandlmaier,f.galasso}@osram.com
An end-to-end neural network is used for the joint tasks of detection, tracking and identification feature learning. The model takes two consecutive video frames as input, and yields: (1) detections in each frame, along with (2) the identification feature of each detection, and (3) predictions from the first frame to next. No separate networks are used for the detection and identification feature learning. Cosine similarity of the identity feature and the IoU of detections are used by the tracking algorithm for detection association.

B.2. Median Flow detection enrichment for Vehicle Tracking (MFOMOT)

Dinh-Quoc-Trung Dang, Vu-Gia-Hy Che, Tien Dinh
{ddqtrung, cvghy}@apcs.vn, dbtien@fit.hcmus.edu.vn
MFOMOT is a simple extension to the traditional tracking-by-detection paradigm. A single object tracker is used to obtain additional detection hypotheses [9], while keeping the whole tracking process online. This method combines both the shape and color features to extract information about the appearance of targets. Visual cues help both in the target association and the elimination of potential drifting in the Median Flow tracker.

B.3. Visual Intersection-over-union tracker (V-IOU)

Erik Bochinski, Volker Eiselein, Thomas Sikora
{bochinski,eiselein,sikora}@nue.tu-berlin.de
V-IOU is based on the IOU tracker [3] and improved by track continuation using the 'visual tracker' if no detection is available. If a valid detection can be associated to the track again, the 'visual tracking' is stopped and the tracker reduces to the original IOU tracker. Otherwise, the visual tracking is aborted after t frames. For each new track, the visual tracking is performed backwards for a maximum of t previous frames or until the track can be merged with a finished track, if the IOU criteria of [3] is satisfied. This extension is made to efficiently reduce the high amount of fragmentation of the tracks produced by the original IOU tracker. Commonly, the visual trackers can not reliably determine if a track was lost or if it should end. For multiple object tracking problems, tracks start and end continuously as objects enter and leave the scene. This continuation of each track by visual tracking would therefore produce a high amount of false positive stubs at the end of each track where visual tracking can not be performed properly. Therefore, each track is required to start and end with an input detection. Thus all visually tracked bounding boxes are removed before reporting to the final tracks.

B.4. Kalman position tracker using IoU-matched detections (KF-IOU)

Oliver Acatay, Lars Sommer, Arne Schumann
{oliver.acatay,lars.sommer,arne.schumann}@iosb.fraunhofer.de
KF-IOU combines the IOU-tracker [3] with a Kalman Filter (KF), with an aim to reduce the number of fragmented tracks and ID switches, as well as to decrease the number of false positives and false negatives. In the first frame, a track is initialized for each detection. For each track, the coordinates of the detections are used to initialize the KF state. In the next frame, KF performs a prediction step. The estimated coordinates in combination with the height and width of the previous detection are used to search for a matching bounding box among all detections in the frame. The best match is the detection with the highest IoU with the estimated bounding box. If the IoU is above a manually defined threshold, the detection is used to update the KF of the track. Otherwise, the estimated bounding box is used to continue the track. The remaining detections of the frame are again used to initialize new tracks. A track is terminated if no matching detections are found in five consecutive frames. The detections are generated using RefineDet [24]. Two variants of the RefineDet detector are trained: one with SEResNeXt-50 [10] and one with ResNet-50 as base network, each with the same set of anchor sizes. Results from the two detectors are combined via averaging.

B.5. Detection Mean Confidence (DMC)

Young Chul Lim, Minsung Kang
{ninolyc,mskang}@dgist.ac.kr
DMC is based on [11], where multi-channel feature generation, pedestrian detection, visual tracking, and data association are all combined to increase the computational efficiency through the sharing of multi-channel features. Multi-channel features (three color channels and seven gradient channels) are generated from a given input image. In the object detector, feature vectors are established by aggregating the multi-channel features. The visual tracker operates on the color channel and gradient channel images using a multi-channel kernelized correlation filter scheme [10]. Hungarian algorithm-based data association then assigns the detections to the tracks, by calculating the similarity costs based on: (1) a histogram-based appearance model and (2) the spatial overlapping between the detections and tracks. Unlike an earlier method [11], this approach manages various track states such as track null, track initialization, track activation, and track termination states using (1) the mean confidence of consecutive detections and (2) the minimum track length.

B.6. Online and Real-Time Tracking with the GM-PHD Filter using Group Management and Relative Motion Analysis (GMMA)

Young-min Song, Moongu Jeon
{kutschbach,eiselein}@nue.tu-berlin.de

GMMA is an online multiple object tracking (MOT) framework, which includes a two-stage data association strategy with the Gaussian mixture probability hypothesis density (GM-PHD) filter [20]. GMMA also includes an occlusion handling method based on group management and motion analysis. The two-stage data association aims to solve the inherent limitations of online tracking. In first stage, a tracker initializes the targets' states from the initial detections. Then, the tracking results are stacked into a set of tracklets. However, because online process inherently is not able to understand the track assignment in view of global optimization, the conventional GM-PHD filter [20] is extended to handle online MOT.

B.7. Semantic Color Multi-object Tracker in Three Level Data Association (SCTrack-3L)

Noor M. Al-Shakarji, Filiz Bunyak, Guna Seetharaman, Kannappan Palaniappan
{nmahyd,bunyak}@missouri.edu,
gunasekaran.seetharaman@rl.af.mil, palaniappank@missouri.edu

SCTrack-3L is based on [1] with an extension to include multiple level of data association. This method includes a time-efficient detection-based multi-object tracking system based on a three-component cascaded data association scheme. Specifically, the three components include: (1) a fast spatial-distance-only short-term data association, (2) a robust tracklet linking step using discriminative object appearance models, and (3) an explicit occlusions handling unit, which relies on both motion patterns and environmental constraints (such as presence of potential occluders) in the scene.

References

[1] N. M. Al-Shakarji, G. Seetharaman, F. Bunyak, and K. Palaniappan. Robust multi-object tracking with semantic color correlation. In *AVSS*, pages 1–7, 2017.

[2] S. Amin and F. Galasso. Geometric proposals for faster R-CNN. In *AVSS*, pages 1–6, 2017.

[3] E. Bochinski, V. Eiselein, and T. Sikora. High-speed tracking-by-detection without using image information. In *AVSS*, pages 1–6, 2017.

[4] Z. Cai, M. J. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, pages 3361–3369, 2015.

[5] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016.

[6] P. Dollár, R. Appel, S. J. Belongie, and P. Perona. Fast feature pyramids for object detection. *TPAMI*, 36(8):1532–1545, 2014.

[7] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[8] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[9] Q. He, J. Wu, G. Yu, and C. Zhang. SOT for MOT. *CoRR*, abs/1712.01059, 2017.

[10] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017.

[11] Y. Lim and M. Kang. Multi-pedestrian detection and tracking using unified multi-channel features. In *AVSS*, pages 1–5, 2017.

[12] Y.-C. Lim and M. Kang. Object detection using a single extended feature map. *IEEE Intelligent Vehicles Symposium*, 2018.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[14] S. Lyu, M. Chang, D. Du, L. Wen, H. Qi, Y. Li, Y. Wei, L. Ke, T. Hu, M. D. Coco, P. Carcagnì, and *et al.* UA-DETRAC 2017: Report of AVSS2017 & IWT4S challenge on advanced traffic monitoring. In *AVSS*, pages 1–7, 2017.

[15] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, pages 1201–1208, 2011.

[16] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[17] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[18] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2017.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[20] B.-N. Vo and W.-K. Ma. The Gaussian mixture probability hypothesis density filter. *TSP*, 54(11):4091, 2006.

[21] L. Wang, Y. Lu, H. Wang, Y. Zheng, H. Ye, and X. Xue. Evolving boxes for fast vehicle detection. In *ICME*, pages 1135–1140, 2017.

[22] L. Wen, D. Du, Z. Cai, Z. Lei, M. Chang, H. Qi, J. Lim, M. Yang, and S. Lyu. UA-DETRAC: A new benchmark and protocol for multi-object tracking. *CoRR*, abs/1511.04136, 2015.

[23] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. *CoRR*, abs/1703.07402, 2017.

[24] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018.

[25] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. *CoRR*, abs/1707.07012, 2017.