Multiple Object Tracking with Drones

Youngbin Ro
2019020546
Korea University
youngbin_ro@korea.ac.kr

1. Introduction

Multiple Object Tracking (MOT) has made many advances with the remarkable development of detection algorithms such as Faster-RCNN[14] and YOLO[13]. However, the assignment problem[4] after detecting objects in the given frames is still remained as the challenging problem in this research area. Especially with videos from drones, the task is much harder because view points change and scale faster than usual and it makes the tracker difficult to deal with similar appearances as well as interaction between objects.

In this paper, I used MOTDT[2] which is one of the state-of-the-art algorithms in MOT challenge as a baseline model and revise it to handle videos from drones. A Homography based bounding box addition as well as an assignment scheme combining intersection over union (IoU) and person re-identification (ReID) features are introduced. After the experiment, I derive some implications from the comparison between results of general MOT dataset and Drone dataset.

2. Related Works

2.1. Detection-based Tracking

Multiple Object Tracking has two prominent tracking approaches[10]: Detection-based Tracking (DBT) and Detection-free Tracking (DFT). In recent years, the first one has become the main stream of the area due to advantages of detection algorithms. DBT isn't limited from the number of objects and doesn't need any manual initialization compared to DFT. A main disadvantage of DBT is the performance highly depends on that of detectors, which is addressed a lot. For example, Bochinski *et al*.[1] used Mask R-CNN[7] as a base detector and achieved state-of-the-art performance on VisDrone Challenge[12]. Feichtenhofer *et al*.[5] fine-tune R-FCN[3] to make the detector track objects simultaneously.

As a result, Many researchers are now focusing on how to assign detections to existing tracks rather than improving performances of detectors. I also used given detection bounding boxes from MOT Challenge[11] and VisDrone datasets[12] to conduct the experiment.

2.2. IoU based Assignment

IoU is computed as area of intersection divided over area of union between bounding boxes in object detection and tracking. It is a main criteria to evaluate the similarity of position from the series of bounding boxes. Many tracking algorithms used the IoU as the assignment scheme because it is reasonable to assume that detections and tracks are the same objects, if they overlap a lot. Chen *et al*.[2] which is my baseline model used IoU as one of the assigning methods in their proposed tracker, and Schulter *et al*.[16] also used IoU in a cost function to evaluate the track assignment. Bochinski *et al*.[1] only used IoU for tracking and they did not even use any image information from frames, which demonstrates IoU is the quite reliable measure in the research area.

the main advantage of using IoU as an assigning policy is the computational efficiency. It doesn't use any forward or backward passes unlike some model-based algorithms. However, IoU suffer from changes in shape of objects and poor performances of detectors more than other measures.

2.3. ReID feature based Assignment

ReID feature is derived from image recognition model such as GoogLeNet[17], ResNet[8], and etc. which all require many computations in training and inference phases. Nevertheless, it can be useful to track reappearing objects from occlusion or lost because it uses image feature itself and memorize distinct properties of each objects. Chen *et al*.[2] also used ReID features from GoogLeNet backbone as well as IoU to track objects and Feng *et al*.[6] used the same backbone to get ReID features and took advantages of it to generate tracklets.

ReID feature is the important cue for tracking, but it require many computation to infer and sometimes vulnerable to distinguish specific objects in crowds. In this paper, I used both IoU and ReID feature with hyper-parameter β to handle ratio of two methods.

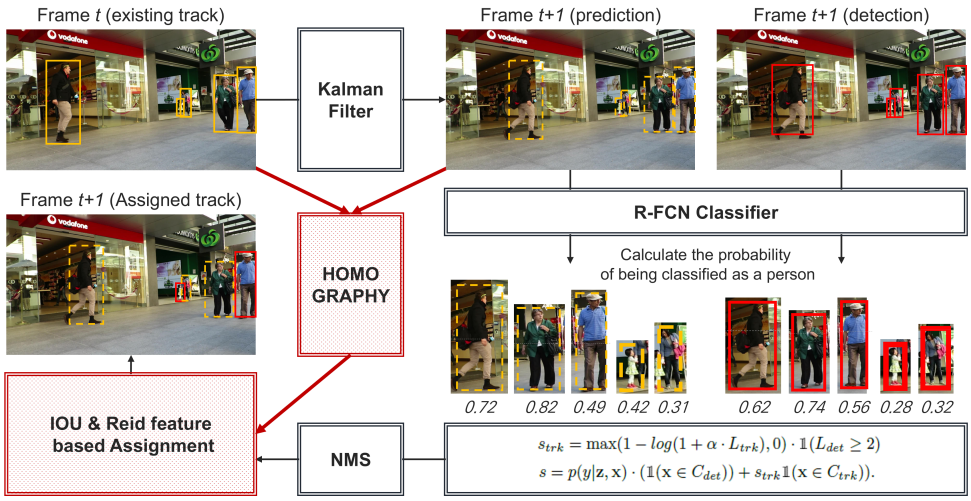

Figure 1. Structure of proposed tracker. The black boxes are from baseline model and the red boxes are proposed methods

3. The Proposed Framework

3.1. Baseline Model

My baseline model is called MOTDT proposed by Chen *et al.*[2]. It has state-of-the-art performance in MOT17 Challenge[11] among real-time tracking models. It only tracks pedestrians and excludes other classes, so I followed same scheme. Its main contribution is unified score function to assign candidate detections to tracks.

In a current frame t, there are existing tracks which have to be assigned with candidate bounding boxes. MOTDT uses Kalman Filter[9] to predict the bounding boxes in the nest frame $t+1$ with original tracks in frame t. The predicted bounding boxes are different from detection bounding boxes which are derived from detectors, and they are all candidates for tracks. After predicting with Kalman Filter, MOTDT apply R-FCN[3] to the candidates to calculate the probability of being classified as a person. With these probabilities, it can get scores with the aforementioned score function below.

$$s_{trk} = \max\left(1 - \log\left(1 + \alpha \cdot L_{trk}\right), 0\right) \cdot \mathbf{1}(L_{det} \leq 2)$$
$$s = p(y \mid z, x) \cdot (\mathbf{1}(x \in C_{det})) + s_{trk}\mathbf{1}(x \in C_{trk})$$

where s_{trk} means a track score which is used for calculating scores for bounding boxes from Kalman Filter, and L_{trk} is the number of track predictions after the last detection is assigned. L_{det} is number of detection results assigned to the track. If the bounding boxes are from detectors, its scores are just the probabilities themselves from R-FCN while the scores for Kalman Filter predictions are multiplication of probabilities and track scores. After the function, Non-Maximum Suppression is conducted and some bounding boxes are filtered out. Lastly, ReID feature based assignment is conducted with reliable candidates while IoU based assignment is used to assign remaining candidates after assigning with ReID features.

3.2. Homography-based Prediction

The first proposal is to generate the supplementary bounding boxes with Homography matrix derived from consecutive two frames t and $t+1$. Firstly, the ORB features[15] and key-point pairs from the features are extracted from the frames. The Homography matrix H can be solved in the closed form with the point pairs and the bounding box vector b is perspectively transformed with the Homography matrix as $b^T H$. This can be done with all bounding boxes in the frame t. As a result, the additional bounding boxes which take a role of supplementary candidates after assigning with original method are proposed.

3.3. New Assignment Scheme

Originally all candidates are firstly assigned with ReID features and remaining features are assigned only with IoU. It makes the tracker fast, but can be a naive approach in that it ignores the image information. The proposed tracker use both methods and hyper-parameter β to handle the ratio of IoU and ReID features' effects. The β is weight for IoU assignment while $1 - \beta$ means weight for the ReID features for assignment. The β for drone dataset is set as 0.8 because image patches in the dataset are too small to consider the ReID features as reliable properties. On the contrary, the β for general MOT dataset is set as 0.2 because ReID features are reliable.

MOT17	IDF1	MT	ML	FP	FN	IDs	Frag.	MOTA	MOTP
Original	0.503	59	**151**	**919**	28580	200	706	**0.428**	0.152
Proposed	**0.522**	**70**	152	3057	**26781**	**198**	**574**	0.421	**0.164**

VisDrone2019	IDF1	MT	ML	FP	FN	IDs	Frag.	MOTA	MOTP
Original	0.547	75	**94**	**725**	22704	504	1604	0.524	0.094
Proposed	**0.579**	**97**	97	3064	**19818**	**386**	**806**	**0.538**	**0.116**

Table 1. MOT17 (above) and VisDrone2019 (below) Results. The better performances are shown in bold.

4. Experiment

The experiment is conducted with both original MOTDT and proposed method. The evaluation and comparison are done after running models. The used datasets and evaluation measures are introduced below.

4.1. Datasets

For general MOT, MOT17 Challenge dataset is used. It consists of 7 clips on both training and testing set and each has 5316 and 5919 frames. VisDrone2019 dataset is used for drone dataset. It consists of 56 clips for training, 7 for validation, and 33 for testing. Each has 24201, 2819, and 12968 frames. The all bounding boxes are provided from original datasets and no additional detectors are not introduced. The Faster R-CNN are used to generate the bounding boxes according to the challenge.

4.2. Evaluation Measures

I followed the evaluation protocol of MOT16[11] to evaluate the performance of proposed tracker. The MOTA metric combines three error sources: FP, FN, and IDs. The MOTP metric is the average dissimilarity between all true positives and the corresponding ground truth targets. IDF1 is F1 score of tracked IDs, which is the ratio of correctly identified detections over the average number of ground truth and computed detections. MT and ML are also important measure in this area. MT is mostly tracked targets, which means the ratio of ground truth trajectories that are covered by a tracker for at least 80% of their respective life span. On the other hand, ML which means mostly lost is the ratio of ground truth trajectories that are covered by a tracker for at most 20% of their respective life span. FP and FN are false positives and false negatives. IDs means the total number of identity switches while Fragment is the total number of times a trajectory is fragmented.

4.3. Results

There is clear trade-off between the original and proposed method. Results are shown in Table1. FP increased a lot in proposed method compared to original one because of the additional bounding boxes generated by Homography, while MT measure is improved. This demonstrates that the additional bounding boxes are helpful to track lost objects in original one, but there are some unnecessary boxes added by Homography. In addition to improvements on MT measure, FN and Fragment measure decreased considerably due to the supplementary bounding boxes.

Tracking time which is not shown in the table increased enormously because extracting ORB features and deriving Homography matrix needs hard computation procedure. The additional usage of ReID feature in the assignment phase also contributed to the increase of tracking time.

In non-numerical evaluation by analyzing video results, both improvements and drawbacks are all observed. The proposed method can capture the same objects even though the perspective of cameras are changed while the original one can not. However, some floating bounding boxes are observed also in the proposed one which increased the FP. The proposed one is also vulnerable to track objects in crowds of drone dataset. It is estimated that the usage of ReID feature in the dataset makes the tracker confused more than that of original one.

In addition to the analysis of each dataset, comparison between two datasets are also conducted. The proposed method performs well relatively in VidDrone2019 in that MOTA is improved in the drone dataset while the other one is not. The usage of β as 0.8 in drone dataset and 0.2 in general MOT dataset is also proper. This indicates that IoU is more reliable than ReID feature in drone dataset.

5. Conclusion

In this paper, I revised the baseline model MOTDT with Homography-based bounding boxes. More strict assignment scheme is also introduced in the proposed method with hyper-parameter β. The proposed one doesn't guarantee the superior results on MOT task, but there are improvements on some measures such as IDF1, MT, FN, IDs, Fragment, and MOTP. However, increases of ML and FP measure due to the additional bounding boxes are main drawbacks of proposed method. Tracking time also can be the severe problem for real-time tracking. Nonetheless, the study get implications that the IoU measure is more reliable property than ReID feature in drone dataset. Additionally, other candidate addition methods can be proposed in the future study.

References

[1] E. Bochinski, T. Senst, and T. Sikora. Extending iou based multi-object tracking by visual information. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

[2] L. Chen, H. Ai, Z. Zhuang, and C. Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2018.

[3] J. Dai, Y. L., K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *NIPS*, 2016.

[4] P. Emami, P. M. Pardalos, L. Elefteriadou, and S. Ranka. Machine learning methods for solving assignment problems in multi-target tracking. *CoRR*.

[5] C. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to track and track to detect. *IEEE International Conference on Computer Vision*, 2017.

[6] W. Feng, Z. Hu, W. Wu, J. Yan, and W. Ouyang. Multi-object tracking with multiple cues and switcher-aware classification. *CoRR*, pages 770–778, 2019.

[7] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82:35–45, 1960.

[10] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, and T. Kim. Multiple object tracking: A literature review. 2014.

[11] A. Milan, L. Leal-Taixé, T. D. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *CoRR*, 2016.

[12] Z. Pengfei, W. Longyin, B. Xiao, H. Ling, and H. Qinghua. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.

[13] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[14] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems*, 1:91–99, 2015.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. Orb: An efficient alternative to sift or surf. *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.

[16] S. Schulter, P. Vernaza, W. Choi, and M. K. Chandraker. Deep network flow for multi-object tracking. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2017.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.