SIMPLE ONLINE AND REALTIME TRACKING

Alex Bewley[†], *Zongyuan Ge*[†], *Lionel Ott*[◇], *Fabio Ramos*[◇], *Ben Upcroft*[†]

Queensland University of Technology[†], University of Sydney[◇]

ABSTRACT

This paper explores a pragmatic approach to multiple object tracking where the main focus is to associate objects efficiently for online and realtime applications. To this end, detection quality is identified as a key factor influencing tracking performance, where changing the detector can improve tracking by up to 18.9%. Despite only using a rudimentary combination of familiar techniques such as the Kalman Filter and Hungarian algorithm for the tracking components, this approach achieves an accuracy comparable to state-of-the-art online trackers. Furthermore, due to the simplicity of our tracking method, the tracker updates at a rate of 260 Hz which is over 20x faster than other state-of-the-art trackers.

Index Terms— Computer Vision, Multiple Object Tracking, Detection, Data Association

1. INTRODUCTION

This paper presents a lean implementation of a tracking-by-detection framework for the problem of multiple object tracking (MOT) where objects are detected each frame and represented as bounding boxes. In contrast to many batch based tracking approaches [1, 2, 3], this work is primarily targeted towards online tracking where only detections from the previous and the current frame are presented to the tracker. Additionally, a strong emphasis is placed on efficiency for facilitating realtime tracking and to promote greater uptake in applications such as pedestrian tracking for autonomous vehicles.

The MOT problem can be viewed as a data association problem where the aim is to associate detections across frames in a video sequence. To aid the data association process, trackers use various methods for modelling the motion [1, 4] and appearance [5, 3] of objects in the scene. The methods employed by this paper were motivated through observations made on a recently established visual MOT benchmark [6]. Firstly, there is a resurgence of mature data association techniques including Multiple Hypothesis Tracking (MHT) [7, 3] and Joint Probabilistic Data Association (JPDA) [2] which occupy many of the top positions of the MOT benchmark. Secondly, the only tracker that does not use the Aggregate Channel Filter (ACF) [8] detector is also

Thanks to ACARP for funding.

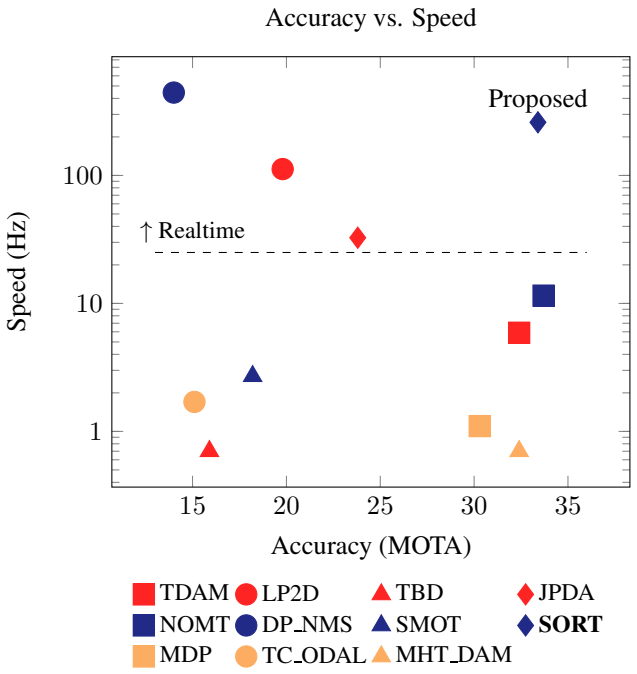

Fig. 1. Benchmark performance of the proposed method (**SORT**) in relation to several baseline trackers [6]. Each marker indicates a trackers accuracy and speed measured in frames per second (FPS) [Hz], i.e. higher and more right is better.

the top ranked tracker, suggesting that detection quality could be holding back the other trackers. Furthermore, the trade-off between accuracy and speed appears quite pronounced, since the speed of most accurate trackers is considered too slow for realtime applications (see Fig. 1). With the prominence of traditional data association techniques among the top online and batch trackers along with the use of different detections used by the top tracker, this work explores how simple MOT can be and how well it can perform.

Keeping in line with Occam's Razor, appearance features beyond the detection component are ignored in tracking and only the bounding box position and size are used for both motion estimation and data association. Furthermore, issues regarding short-term and long-term occlusion are also ignored, as they occur very rarely and their explicit treatment intro-

duces undesirable complexity into the tracking framework. We argue that incorporating complexity in the form of object re-identification adds significant overhead into the tracking framework – potentially limiting its use in realtime applications.

This design philosophy is in contrast to many proposed visual trackers that incorporate a myriad of components to handle various edge cases and detection errors [9, 10, 11, 12]. This work instead focuses on efficient and reliable handling of the common frame-to-frame associations. Rather than aiming to be robust to detection errors, we instead exploit recent advances in visual object detection to solve the detection problem directly. This is demonstrated by comparing the common ACF pedestrian detector [8] with a recent convolutional neural network (CNN) based detector [13]. Additionally, two classical yet extremely efficient methods, Kalman filter [14] and Hungarian method [15], are employed to handle the motion prediction and data association components of the tracking problem respectively. This minimalistic formulation of tracking facilitates both efficiency and reliability for online tracking, see Fig. 1. In this paper, this approach is only applied to tracking pedestrians in various environments, however due to the flexibility of CNN based detectors [13], it naturally can be generalized to other objects classes.

The main contributions of this paper are:

- We leverage the power of CNN based detection in the context of MOT.
- A pragmatic tracking approach based on the Kalman filter and the Hungarian algorithm is presented and evaluated on a recent MOT benchmark.
- Code will be open sourced to help establish a baseline method for research experimentation and uptake in collision avoidance applications.

This paper is organised as follows: Section 2 provides a short review of related literature in the area of multiple object tracking. Section 3 describes the proposed lean tracking framework before the effectiveness of the proposed framework on standard benchmark sequences is demonstrated in Section 4. Finally, Section 5 provides a summary of the learnt outcomes and discusses future improvements.

2. LITERATURE REVIEW

Traditionally MOT has been solved using Multiple Hypothesis Tracking (MHT) [7] or the Joint Probabilistic Data Association (JPDA) filters [16, 2], which delay making difficult decisions while there is high uncertainty over the object assignments. The combinatorial complexity of these approaches is exponential in the number of tracked objects making them impractical for realtime applications in highly dynamic environments. Recently, Rezatofighi et al. [2], revisited the JPDA formulation [16] in visual MOT with the goal to address the combinatorial complexity issue with an efficient approximation of the JPDA by exploiting recent developments in solving integer programs. Similarly, Kim et al. [3] used an appearance model for each target to prune the MHT graph to achieve state-of-the-art performance. However, these methods still delay the decision making which makes them unsuitable for online tracking.

Many online tracking methods aim to build appearance models of either the individual objects themselves [17, 18, 12] or a global model [19, 11, 4, 5] through online learning. In addition to appearance models, motion is often incorporated to assist associating detections to tracklets [1, 19, 4, 11]. When considering only one-to-one correspondences modelled as bipartite graph matching, globally optimal solutions such as the Hungarian algorithm [15] can be used [10, 20].

The method by Geiger et al. [20] uses the Hungarian algorithm [15] in a two stage process. First, tracklets are formed by associating detections across adjacent frames where both geometry and appearance cues are combined to form the affinity matrix. Then, the tracklets are associated to each other to bridge broken trajectories caused by occlusion, again using both geometry and appearance cues. This two step association method restricts this approach to batch computation. Our approach is inspired by the tracking component of [20], however we simplify the association to a single stage with basic cues as described in the next section.

3. METHODOLOGY

The proposed method is described by the key components of detection, propagating object states into future frames, associating current detections with existing objects, and managing the lifespan of tracked objects.

3.1. Detection

To capitalise on the rapid advancement of CNN based detection, we utilise the Faster Region CNN (**FrRCNN**) detection framework [13]. **FrRCNN** is an end-to-end framework that consists of two stages. The first stage extracts features and proposes regions for the second stage which then classifies the object in the proposed region. The advantage of this framework is that parameters are shared between the two stages creating an efficient framework for detection. Additionally, the network architecture itself can be swapped to any design which enables rapid experimentation of different architectures to improve the detection performance.

Here we compare two network architectures provided with **FrRCNN**, namely the architecture of Zeiler and Fergus (**FrRCNN(ZF)**) [21] and the deeper architecture of Simonyan and Zisserman (**FrRCNN(VGG16)**) [22]. Throughout this work, we apply the **FrRCNN** with default parameters learnt for the PASCAL VOC challenge. As we are only interested in pedestrians we ignore all other classes and only pass person detection results with output probabilities greater than 50% to the tracking framework.

Table 1. Comparison of tracking performance by switching the detector component. Evaluated on Validation sequences as listed in [12].

Tracker	Detector	Detection		Tracking	
		Recall	Precision	ID Sw	MOTA
MDP [12]	ACF	36.6	75.8	222	24.0
	FrRCNN(ZF)	46.2	67.2	245	22.6
	FrRCNN(VGG16)	**50.1**	76.0	**178**	33.5
Proposed	ACF	33.6	65.7	224	15.1
	FrRCNN(ZF)	41.3	72.4	347	24.0
	FrRCNN(VGG16)	49.5	**77.5**	274	**34.0**

In our experiments, we found that the detection quality has a significant impact on tracking performance when comparing the **FrRCNN** detections to **ACF** detections. This is demonstrated using a validation set of sequences applied to both an existing online tracker **MDP** [12] and the tracker proposed here. Table 1 shows that the best detector (**FrRCNN(VGG16)**) leads to the best tracking accuracy for both **MDP** and the proposed method.

3.2. Estimation Model

Here we describe the object model, i.e. the representation and the motion model used to propagate a target's identity into the next frame. We approximate the inter-frame displacements of each object with a linear constant velocity model which is independent of other objects and camera motion. The state of each target is modelled as:

$$\mathbf{x} = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T,$$

where u and v represent the horizontal and vertical pixel location of the centre of the target, while the scale s and r represent the scale (area) and the aspect ratio of the target's bounding box respectively. Note that the aspect ratio is considered to be constant. When a detection is associated to a target, the detected bounding box is used to update the target state where the velocity components are solved optimally via a Kalman filter framework [14]. If no detection is associated to the target, its state is simply predicted without correction using the linear velocity model.

3.3. Data Association

In assigning detections to existing targets, each target's bounding box geometry is estimated by predicting its new location in the current frame. The assignment cost matrix is then computed as the intersection-over-union (IOU) distance between each detection and all predicted bounding boxes from the existing targets. The assignment is solved optimally using the Hungarian algorithm. Additionally, a minimum IOU is imposed to reject assignments where the detection to target overlap is less than IOU_{min}.

We found that the IOU distance of the bounding boxes implicitly handles short term occlusion caused by passing targets. Specifically, when a target is covered by an occluding object, only the occluder is detected, since the IOU distance appropriately favours detections with similar scale. This allows both the occluder target to be corrected with the detection while the covered target is unaffected as no assignment is made.

3.4. Creation and Deletion of Track Identities

When objects enter and leave the image, unique identities need to be created or destroyed accordingly. For creating trackers, we consider any detection with an overlap less than IOU_{min} to signify the existence of an untracked object. The tracker is initialised using the geometry of the bounding box with the velocity set to zero. Since the velocity is unobserved at this point the covariance of the velocity component is initialised with large values, reflecting this uncertainty. Additionally, the new tracker then undergoes a probationary period where the target needs to be associated with detections to accumulate enough evidence in order to prevent tracking of false positives.

Tracks are terminated if they are not detected for T_{Lost} frames. This prevents an unbounded growth in the number of trackers and localisation errors caused by predictions over long durations without corrections from the detector. In all experiments T_{Lost} is set to 1 for two reasons. Firstly, the constant velocity model is a poor predictor of the true dynamics and secondly we are primarily concerned with frame-to-frame tracking where object re-identification is beyond the scope of this work. Additionally, early deletion of lost targets aids efficiency. Should an object reappear, tracking will implicitly resume under a new identity.

4. EXPERIMENTS

We evaluate the performance of our tracking implementation on a diverse set of testing sequences as set by the MOT benchmark database [6] which contains both moving and static camera sequences. For tuning the initial Kalman filter covariances, IOU_{min}, and T_{Lost} parameters, we use the same training/validation split as reported in [12]. The detection architecture used is the **FrRCNN(VGG16)** [22]. Source code and sample detections from [22] are available online. [1]

4.1. Metrics

Since it is difficult to use one single score to evaluate multi-target tracking performance, we utilise the evaluation metrics defined in [24], along with the standard MOT metrics [25]:

- MOTA(↑): Multi-object tracking accuracy [25].
- MOTP(↑): Multi-object tracking precision [25].

[1] https://github.com/abewley/sort

Table 2. Performance of the proposed approach on MOT benchmark sequences [6].

Method	Type	MOTA↑	MOTP↑	FAF↓	MT↑	ML↓	FP↓	FN↓	ID sw↓	Frag↓
TBD [20]	Batch	15.9	70.9	2.6%	6.4%	47.9%	14943	34777	1939	1963
ALExTRAC [5]	Batch	17.0	71.2	1.6%	3.9%	52.4%	9233	39933	1859	1872
DP_NMS [23]	Batch	14.5	70.8	2.3%	6.0%	40.8%	13171	34814	4537	3090
SMOT [1]	Batch	18.2	71.2	1.5%	2.8%	54.8%	8780	40310	1148	2132
NOMT [11]	Batch	**33.7**	71.9	**1.3%**	12.2%	44.0%	7762	32547	**442**	**823**
RMOT [4]	**Online**	18.6	69.6	2.2%	5.3%	53.3%	12473	36835	684	1282
TC_ODAL [17]	**Online**	15.1	70.5	2.2%	3.2%	55.8%	12970	38538	637	1716
TDAM [18]	**Online**	33.0	**72.8**	1.7%	**13.3%**	39.1%	10064	**30617**	464	1506
MDP [12]	**Online**	30.3	71.3	1.7%	13.0%	38.4%	9717	32422	680	1500
SORT (Proposed)	**Online**	**33.4**	72.1	**1.3%**	11.7%	**30.9%**	**7318**	32615	1001	1764

- FAF(↓): number of false alarms per frame.
- MT(↑): number of mostly tracked trajectories. I.e. target has the same label for at least 80% of its life span.
- ML(↓): number of mostly lost trajectories. i.e. target is not tracked for at least 20% of its life span.
- FP(↓): number of false detections.
- FN(↓): number of missed detections.
- ID sw(↓): number of times an ID switches to a different previously tracked object [24].
- Frag(↓): number of fragmentations where a track is interrupted by miss detection.

Evaluation measures with (↑), higher scores denote better performance; while for evaluation measures with (↓), lower scores denote better performance. True positives are considered to have at least 50% overlap with the corresponding ground truth bounding box. Evaluation codes were downloaded from [6].

4.2. Performance Evaluation

Tracking performance is evaluated using the MOT benchmark [6] test server where the ground truth for 11 sequences is withheld. Table 2 compares the proposed method **SORT** with several other baseline trackers. For brevity, only the most relevant trackers, which are state-of-the-art online trackers in terms of accuracy, such as (**TDAM** [18], **MDP** [12]), the fastest batch based tracker (**DP_NMS** [23]), and all round *near* online method (**NOMT** [11]) are listed. Additionally, methods which inspired this approach (**TBD** [20], **ALExTRAC** [5], and **SMOT** [1]) are also listed. Compared to these other methods, **SORT** achieves the highest MOTA score for the online trackers and is comparable to the state-of-the-art method **NOMT** which is significantly more complex and uses frames in the near future. Additionally, as **SORT** aims to focus on frame-to-frame associations the number of lost targets (**ML**) is minimal despite having similar false negatives to other trackers. Furthermore, since **SORT** focuses on frame-to-frame associations to grow tracklets, it has the lowest number of lost targets in comparison to the other methods.

4.3. Runtime

Most MOT solutions aim to push performance towards greater accuracy, often, at the cost of runtime performance. While slow runtime may be tolerated in offline processing tasks, for robotics and autonomous vehicles, realtime performance is essential. Fig. 1 shows a number of trackers on the MOT benchmark [6] in relation to both their speed and accuracy. This shows that methods which achieve the best accuracy also tend to be the slowest (bottom right in Figure 1). On the opposite end of the spectrum the fastest methods tend to have lower accuracy (top left corner in Figure 1). **SORT** combines the two desirable properties, speed and accuracy, without the typical drawbacks (top right in Figure 1). The tracking component runs at 260 Hz on single core of an Intel i7 2.5GHz machine with 16 GB memory.

5. CONCLUSION

In this paper, a simple online tracking framework is presented that focuses on frame-to-frame prediction and association. We showed that the tracking quality is highly dependent on detection performance and by capitalising on recent developments in detection, state-of-the-art tracking quality can be achieved with only classical tracking methods. The presented framework achieves best in class performance with respect to both speed and accuracy, while other methods typically sacrifice one for the other. The presented framework's simplicity makes it well suited as a baseline, allowing for new methods to focus on object re-identification to handle long term occlusion. As our experiments highlight the importance of detection quality in tracking, future work will investigate a tightly coupled detection and tracking framework.

6. REFERENCES

[1] C. Dicle, M. Sznaier, and O. Camps, "The way they move: Tracking multiple targets with similar appearance," in *International Conference on Computer Vision*, 2013.

[2] S. H. Rezatofighi, A. Milan, Z. Zhang, A. Dick, Q. Shi, and I. Reid, "Joint Probabilistic Data Association Revisited," in *International Conference on Computer Vision*, 2015.

[3] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple Hypothesis Tracking Revisited," in *International Conference on Computer Vision*, 2015.

[4] J. H. Yoon, M. H. Yang, J. Lim, and K. J. Yoon, "Bayesian Multi-Object Tracking Using Motion Context from Multiple Objects," in *Winter Conference on Applications of Computer Vision*, 2015.

[5] A. Bewley, L. Ott, F. Ramos, and B. Upcroft, "ALExTRAC: Affinity Learning by Exploring Temporal Reinforcement within Association Chains," in *International Conference on Robotics and Automation*. 2016, IEEE.

[6] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking," *arXiv preprint*, 2015.

[7] D. Reid, "An Algorithm for Tracking Multiple Targets," *Automatic Control*, vol. 24, pp. 843–854, 1979.

[8] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection," *Pattern Analysis and Machine Intelligence*, vol. 36, 2014.

[9] S. Oh, S. Russell, and S. Sastry, "Markov Chain Monte Carlo Data Association for General Multiple-Target Tracking Problems," in *Decision and Control*. 2004, pp. 735–742, IEEE.

[10] A. Perera, C. Srinivas, A. Hoogs, and G. Brooksby, "Multi-Object Tracking Through Simultaneous Long Occlusions and Split-Merge Conditions," in *Computer Vision and Pattern Recognition*. 2006, IEEE.

[11] W. Choi, "Near-Online Multi-target Tracking with Aggregated Local Flow Descriptor," in *International Conference on Computer Vision*, 2015.

[12] Y. Xiang, A. Alahi, and S. Savarese, "Learning to Track : Online Multi-Object Tracking by Decision Making," in *International Conference on Computer Vision*, 2015.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, 2015.

[14] R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[15] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[16] Y. Bar-Shalom, *Tracking and data association*, Academic Press Professional, Inc., 1987.

[17] S. H. Bae and K. J. Yoon, "Robust Online Multi-Object Tracking based on Tracklet Confidence and Online Discriminative Appearance Learning," *Computer Vision and Pattern Recognition*, 2014.

[18] Y. Min and J. Yunde, "Temporal Dynamic Appearance Modeling for Online Multi-Person Tracking," oct 2015.

[19] A. Bewley, V. Guizilini, F. Ramos, and B. Upcroft, "Online Self-Supervised Multi-Instance Segmentation of Dynamic Objects," in *International Conference on Robotics and Automation*. 2014, IEEE.

[20] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3D Traffic Scene Understanding from Movable Platforms," *Pattern Analysis and Machine Intelligence*, 2014.

[21] M. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *European Conference on Computer Vision*, 2014.

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, 2015.

[23] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Computer Vision and Pattern Recognition*. 2011, IEEE.

[24] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: HybridBoosted multi-target tracker for crowded scene," in *Computer Vision and Pattern Recognition*. 2009, IEEE.

[25] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *Image and Video Processing*, , no. May, 2008.