Machine learning based multi-object tracking without dynamic models and hard association metrics

Christian Alexander Holz, Christian Bader, Matthias Drüppel

***Abstract*—In this paper, we develop Machine Learning (ML)-based methods for Multi Object Tracking (MOT) within the context of Advanced Driver Assistance Systems (ADAS). Given the increasing complexity and demand for precise and efficient object tracking systems in the automotive industry, this work focuses on the integration of ML techniques into established tracking methodologies. Key contributions encompass the creation and evaluation of three specialized neural networks: (i) the Single Prediction Network (SPENT) for predicting the trajectories of tracked objects, (ii) the Single Association Network (SANT) for associating incoming sensor objects with existing tracks, (iii) and the Multi Association Network (MANTa) for associating multiple sensor objects with existing tracks. Figure 1 provides an overview of our approach (i) and (ii). These networks aim to combine ML methods with a traditional Kalman filter framework, offering a data driven approach to addressing MOT challenges. We integrate our three ML networks into a Kalman framework and evaluate the performance, both of the components itself and the overall system. By replacing single components, we get a clearer understanding of the impact of the ML models on the overall tracking system. This approach also leaves the modularity of system intact while enabling machine learning (ML) for certain tracking tasks. The results reveal a modular, robust, and maintainable tracker, underscoring the potential of ML integration in ADAS.**

***Index Terms*—Article submission, IEEE, IEEEtran, journal, LaTeX, paper, template, typesetting.**

I. Introduction

THE ongoing evolution of Advanced Driver Assistance Systems (ADAS) has brought the need for precise and reliable Multi Object Tracking (MOT) into the spotlight [1] [2] [3] [4] [5] [6] [7] . In complex and dynamic environments, as encountered in urban traffic, it is crucial to simultaneously and accurately capture the positions and movements of multiple objects. Tracking multiple objects is a key challenge in computer vision (CV).

In the commonly used Tracking-by-Detection (TbD) paradigm, a tracker fuses detected sensor objects (SO) to create consistent object tracks over time. A key challenge within this paradigm is associating incoming measurements SO with their corresponding existing object tracks or initializing new object tracks.

C. Holz is with Daimler Truck AG, Research and Advanced Development, Stuttgart, Germany

C. Bader is with Daimler Truck AG, Research and Advanced Development, Stuttgart, Germany

M. Drüppel is with the Center for Artificial Intelligence, Duale Hochschule Baden-Württemberg (DHBW), Stuttgart, Germany

Offline methods [8] process the entire video material at once in a batch process. However, these methods are unsuitable for most real-time applications, such as ADAS. In such applications, it is crucial to predict the state of objects immediately after new detections. Therefore, most recent approaches for tracking multiple objects rely on online methods that do not depend on future image information. Online methods use various features to estimate the similarity between the recognised objects and the existing tracks. This can be done on the basis of predicted positions or similarities in appearance [2] [3] [4] [5] . While some approaches only consider the most recent recognition corresponding to a track, other methods integrate temporal information into a track history. For example, methods use recurrent neural networks [4] [5] or attention mechanisms [6] [7] to aggregate temporal information.

For state prediction, in many TbD approaches, Kalman filters and their variants have proven to be effective. However, they reach their limits in more complex scenarios, particularly in the presence of non-linear motion patterns and interactions among multiple objects. In this work, we introduce a novel MOT approach that leverages Machine Learning (ML) to overcome these challenges. We specifically focus on the development and implementation of Neural Networks (NN), which can enable more precise and flexible data-driven object tracking without relying on cumbersome heuristics and hyperparameters.

The data association is typically carried out based on similarity scores calculated between the measurements and the existing tracks. These similarity scores may rely on the latest detection or be aggregated from historical detections. As in Mertz et al [5], the aim of this work was to develop a data-based approach that can learn to completely solve the combinatorial Non Deterministic Polynomial Time (NP) hard optimisation problem of data association. Mertz et al [5] use a distance matrix based on the Euclidean distance measure as input data for the developed DA network, thus replacing an association algorithm such as the Hungarian Algorithm (HA). It can be assumed that the Euclidean distance measure was used as the basis for calculating the ground truth (GT) training data (distance matrices) and for the evaluation. However, this is not explicitly stated. It can therefore be argued that this calculation step deprives the network of the opportunity to follow a different association logic or to learn it on the basis of data. In the context of this work, the hypothesis was therefore put forward that a Gated Recurrent Unit (GRU)-based association network can be formed by an undefined

distance measure to increase the assignment of the temporal memory component and thus of the history.

The aim of this work was therefore to develop an association network that is intended to solve the assignment of one or more sensor objects (SO) to an existing number of object tracks without a defined distance measure.

II. Related work

Our primary contribution is the development and evaluation of three NN that we labeled: (i) the Single Prediction Network (SPENT), (ii) the Single Association Network (SANT), and (iii) the Multi Association Network (MANTa). Figure 1 provides an overview of our approach (i) and (ii). The proposed Sinlge Prediction Network (SPENT) processes the sensor objects (SO) per time step, which consists of a state vector and contains information such as object position, orientation and dimension. SPENT predicts a fixed-dimensional state vector for each SO based on the received state vectors per SO. The output predictions from SPENT are used as input to the Single Association Network (SANT) to build target trajectories or object tracks. In the remainder of the following sections, we will explain the proposed modules in detail.

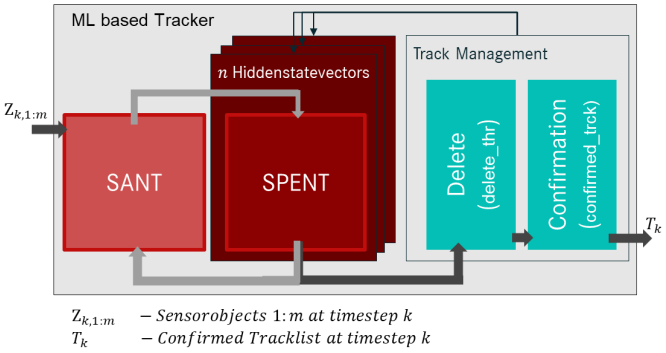

Fig. 1. Schematic representation of the two integrated networks SPENT and SANT in a tracking-by-detection (TbD) framework.

In comparison to the Kalman filter, SPENT is capable of predicting the state of individual objects without the need for a predefined state or observation model at runtime. SPENT holds the potential, particularly in terms of adaptability to various scenarios and the ability to effectively handle nonlinearities. Many conventional tracking systems rely on static methods for data association, often based on simple heuristics or fixed thresholds. In contrast, SANT employs machine learning to automate these processes and adapt more effectively to different scenarios. As a result, within the TbD MOT framework, SANT replaces the calculation of a distance metric and the Hungarian algorithm for the corresponding assignment. Furthermore, we integrate SPENT and SANT into an existing tracking system and demonstrate their performance through multiple tests and comparisons with established methods. This work provides exciting insights and a significant advancement in the development of Advanced Driver Assistance Systems (ADAS), contributing to the further evolution of technologies for autonomous driving (AD).

III. Tracking with Prediction and Association Networks

Wir wenden das Paradigma des Tracking-by-Detection (TbD) an, bei dem ein Tracker die Objekterkennungen fusioniert, um Objektspuren zu erzeugen, die über die Zeit konsistent sind. Ba-Tuong Vo et al. [xx] stellt beispielsweise einen Framework für die Untersuchung von Tracking Ansätzen zur Verfügung, welche dem TbD Paradigma folgen. Wir schlagen einen Tracker vor mit jeweils einem Long Short-Term Memory (LSTM) Netzwerk für die Prädiktion und Assoziation der Sensorobjekte (SO) zu den bestehenden Tracks.

A. Single Prediction Network (SPENT)

Die meisten bestehenden Verfolgungsmethoden verknüpfen eingehende Erkennungen paarweise mit Objektzuständen, die durch ein einfaches Bewegungsmodell, z. B. ein Modell mit konstanter Geschwindigkeit, unter Verwendung eines Kalman-Filters vorhergesagt werden. Neuere Arbeiten haben jedoch gezeigt, dass die Aggregation zeitlicher Informationen sowie von Kontextinformationen die Verfolgung mehrerer Objekte verbessern kann, indem zusätzlich zu den paarweisen Ähnlichkeiten zwischen den Erkennungen Informationen höherer Ordnung genutzt werden [xx]. Unserer Ansatz sieht es vor, auf die Bewegungsmodelle zu verzichten und stattdessen die Hiddenstates der LSTM Schicht als objektspezifisches Parameterset zu nutzen. Die initialen Werte der Hiddenstates der LSTM Schicht werden im Tracking anhand der erhaltenen Messdaten aktualisiert. Durch diese Aktualisierung erfolgt somit eine interne Korrektur über den Sequenzverlauf.

B. Single Association Network (SANT)

C. Multi Association Network (MANTa)

IV. Experimental Evaluation

... KITTI-Car Benchmark.

V. Conclusion

The conclusion goes here.

Acknowledgments

This should be a simple paragraph before the References to thank those individuals and institutions who have supported your work on this article.

Appendix
Proof of the Zonklar Equations

Proof of the First Zonklar Equation

Appendix goes here.

Proof of the Second Zonklar Equation

And here.

REFERENCES

[1] Jenny Seidenschwarz, Guillem Brasó, Victor Serrano, Ismail Elezi, Laura Leal-Taixé, "Simple cues lead to a strong multi-object tracker," *Paper*, 2022.

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, Ben Upcroft, "Simple online and realtime tracking," *Paper*, 2016.

[3] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, Nenghai Yu, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," 2017.

[4] Anton Milan, Seyed Rezatofighi, Anthony Dick, Ian Reid, Konrad Schindler, "Online multi-target tracking using recurrent neural networks," 2016.

[5] H. L. H. Z. C. Mertz, "Deepda: Lstm-based deep data association network for multi-targets tracking in clutter," 2019.

[6] W.-C. H. H. K. T.-Y. L. Y. C. R. Yu, "Soda: Multi-object tracking with soft data association," 2020.

[7] Q. C. W. O. H. L. X. W. B. L. N. Yu, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," 2017.

[8] Roberto Henschel, Laura Leal-Taixé, Daniel Cremers, Bodo Rosenhahn, "Fusion of head and full-body detectors for multi-object tracking," *Paper*, 2017.