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. 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 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]. 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. The challenge here lies not only in the detection and tracking of individual objects but also in considering their interactions and mutual influences, especially in cases of occlusions and sudden changes in motion.

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 (Sensor Objects (SO)) with their corresponding existing tracks or initializing new object tracks. This 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. 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.

Our primary contribution lies in 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). 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 valuable insights and a significant advancement in the development of Advanced Driver Assistance Systems (ADAS), contributing to the further evolution of technologies for autonomous driving.

II. Related work

a) Multi-Object Tracking: Die Verfolgung mehrerer Objekte ist eine zentrale Herausforderung in der Computer Vision. Die meisten bestehenden Ansätze zur Verfolgung mehrerer Objekte, einschließlich des hier vorgestellten, basieren auf dem Ansatz der "Tracking-by-Detection (TbD)". Offline-Methoden [xx, ...] verarbeiten dabei das

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

gesamte Videomaterial auf einmal in einem Stapelverarbeitungsprozess. Diese Methoden sind jedoch für die meisten Echtzeit-Anwendungen, wie beispielsweise ADAS, ungeeignet. In solchen Anwendungen ist es entscheidend, den Zustand von Objekten unmittelbar nach neuen Erkennungen vorherzusagen. Daher setzen die meisten neueren Ansätze zur Verfolgung mehrerer Objekte auf Online-Methoden, die nicht auf zukünftige Bildinformationen angewiesen sind [xx, ...]. Online-Methoden verwenden verschiedene Merkmale, um die Ähnlichkeit zwischen den erkannten Objekten und den existierenden Spuren zu schätzen. Dies kann auf Grundlage von vorhergesagten Positionen oder Ähnlichkeiten im Erscheinungsbild geschehen [xx, ...]. Während einige Ansätze [xx, ...] nur die jüngste Erkennung, die einer Spur entspricht, berücksichtigen, integrieren andere Methoden zeitliche Informationen in eine Spurhistorie. Verfahren nutzen beispielsweise rekurrente neuronale Netze, um zeitliche Informationen zu aggregieren [xx, ...]. Wie von Mertz et al. [xx] wurde auch in dieser Arbeit das Ziel verfolgt einen datenbasierten Ansatz zu entwickeln, welcher lernen kann, das kombinatorische Non Deterministic Polynomial Time (NP) hard Optimierungsproblem der Datenassoziation vollständig zu lösen. Mertz et al. [xx] nutzen als Inputdaten für das entwickelte DA Netzwerk eine Distanzmatrix auf Basis des euklidischen Abstandsmaßes, und ersetzt somit einen Assoziationsalgorithmus wie z.B. den Hungarian Algorithm (HA). Es ist anzunehmen, dass bei der Erstellung der Groundtruth (GT) Trainingsdaten (Distanzmatrizen), sowie bei der Evaluierung, das euklidische Abstandsmaß als Basisberechnung verwendet wurde. Dies ist jedoch nicht explizit aufgeführt. Es lässt sich somit die Behauptung aufstellen, dass durch diesen Berechnungsschritt dem Netzwerk die Möglichkeit genommen wird, einer anderen Assoziationslogik zu folgen bzw. diese datenbasiert zu lernen.

Im Rahmen dieser Arbeit wurde daher die These aufgestellt, dass durch ein nicht definiertes Abstandsmaß ein Gated Recurrent Unit (GRU) basiertes Assoziationsnetzwerk die Zuordnung von der zeitlichen Speicherkomponente und somit von der Historie verstärkt gebildet werden kann. In dieser Arbeit wurde daher das Ziel verfolgt ein Assoziationsnetzwerk zu entwickeln, welches die Zuordnung eines oder mehrerer SO zu einer bestehenden Anzahl an Tracks ohne definiertes Abstandsmaß lösen soll.

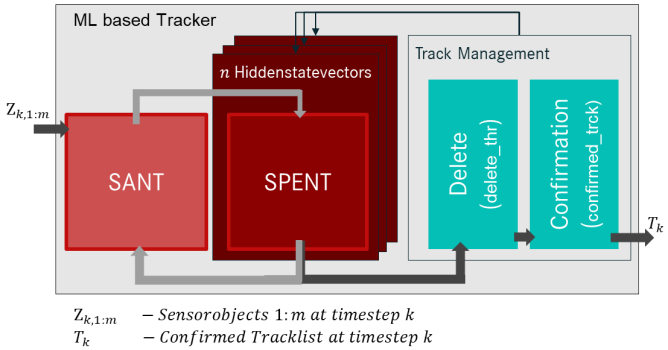

Fig. 1. Schematic representation of the two integrated networks SPENT and SANT in a tracking-by-detection (TbD) framework.

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.

Das vorgeschlagene Sinlge Prediction Network (SPENT) verarbeitet die Erkennungen in einem zeitlichen Fenster, das aus Objektmessungen, wie Position, Objektdimension, relativer Geschwindigkeit und Objekttyp besteht und sagt einen festdimensionalen Zustandsvektor für jede Erkennung voraus. Die ausgegebenen Merkmale dienen dem Single Association Network (SANT) als Input, um Zieltrajektorien zu bilden. Siehe Fig. 1 für einen Überblick über unseren Ansatz. Im weiteren Verlauf dieses Abschnitts werden wir die beiden vorgeschlagenen Module im Detail erläutern.

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.

...

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] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, Ben Upcroft, "Simple online and realtime tracking," *Paper*, 2016.
[2] Anton Milan, Seyed Rezatofighi, Anthony Dick, Ian Reid, Konrad Schindler, "Online multi-target tracking using recurrent neural networks," 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. [Online]. Available: http://arxiv.org/pdf/1708.02843v2
[4] Jenny Seidenschwarz, Guillem Brasó, Victor Serrano, Ismail Elezi, Laura Leal-Taixé, "Simple cues lead to a strong multi-object tracker," 2022.