A collaborative framework for semi-automatic scenario-based mining of big road user data

Imanol Irizar Da Silva[1], Meng Zhang[2] and Kay Gimm[3]

***Abstract*— Traffic research has benefited from a significant expansion in the amount of available data. Consequently, the need arises for an automatic and efficient method to extract and analyze relevant traffic situations instead of a more traditional and manual approach like manual video annotation.**

This paper presents a framework to create such a data pipeline. The user must define the target scenarios and the pipeline will abstract the available trajectory data into candidate scenes (groups of interacting trajectories) and select the matches for the target scenarios. These scenes will be mined and modelled automatically for new valuable information. Furthermore, Surrogate Measures of Safety (SMoS) are applied to identify the critical and atypical scenes of the target scenarios.

A set of eight scenarios containing interactions between bicycles and MRUs (Motorized Road Users) at the AIM (Application Platform for Intelligent Mobility) Research Intersection in the city of Braunschweig, Germany, was mined by a team of three researchers using the presented framework to validate it with positive results.

I. Introduction

In recent years, much more data have become available to researchers. For traffic research, an especially interesting source are trajectory data extracted from videos and other kinds of sensors like LIDAR and radar.

A good example of such a modern, state-of-the art data source is the AIM Research Intersection in Braunschweig, Germany [1]. This platform uses an array of sensors and cameras to detect and track road users on the intersection and generate their real-world trajectories. This live data is streamed and the latest multiple days are kept in a buffer for analysis purpose. The amount of data generated is around 40,000 trajectories per day and they represent more than 5 million potentially interacting pairs. The data is then used by a team of researchers to study a wide array of traffic scenarios and model traffic behaviour. Trajectory data are analysed to get insights into safety critical traffic events and their causes, as well as mechanisms of traffic behaviour. These findings are invaluable for the design of infrastructure, automated driving functions and autonomous vehicles.

The literature contains several kinds of traffic analysis methodologies. First, there are direct observation studies which usually involve a trained observer noting and parameterizing traffic situations using a TCT (Traffic Conflict Technique) [2]. This method is expensive, inconvenient and can work for a small amount of data only. More modern methods use video data to manually find the desired scenarios [2] and even for (semi-)automatic object tracking for the parametrization of these traffic situations [3].

None of these methods are suitable for automatic big data analysis in the context of the AIM Research Intersection for three main reasons:

1) The large amount of data and its limited life in the database requires a scalable and performant data pipeline that is able to run faster than real-time to not lose any data.
2) The big number of scenarios to be mined which share expensive calculations.
3) The variety in the characteristics of the traffic scenarios to be mined which require different mining algorithms.

In short, there is a need for a centralized and scalable big data framework that is able to mine a big number of different scenarios in an automatic, efficient and collaborative manner for several researchers at the same time.

This document is structured as follows. In part II, the state of the art is reviewed in search of methods of data mining, especially in the context of trajectory data and traffic scenarios. In part III, the collaborative scenario mining framework is presented. Part IV shows the results of applying the collaborative framework to the mining of eight scenarios at the AIM Research Intersection. In part V the results are discussed. Part VI is the conclusion.

II. State of the art

The literature reveals a taxonomy and conceptual framework for the analysis of trajectories [4] as well as a great variety of methods and techniques for trajectory data mining [5]. Furthermore, Knowledge Discovery in Databases (KDD) [6] is a well-known and widely used process that consists of 5 steps: selection, pre-processing, transformation, data mining and interpretation/validation. Big Data is a quite popular although not very strictly defined term. "Table I" collects its characteristics [7] and shows that they mostly apply to the data analyzed in the context of this paper.

A very common approach to traffic analysis is to select a traffic scenario, get real data for it and perform a certain analysis on this data [2], [3], [8]. In the context of this paper the term *scenario* refers to an abstract traffic situation while *scene* refers to a specific instance of a scenario. For example, a scenario would be: car turning right at an intersection and

*This work was not supported by any organization

[1]Imanol Irizar Da Silva is a researcher at the Institute of Transportation Systems in DLR (German Aerospace Center), Braunschweig, Germany Imanol.IrizarDaSilva@dlr.de

[2]Meng Zhang is a researcher at the Institute of Transportation Systems in DLR (German Aerospace Center), Berlin, Germany Meng.Zhang@dlr.de

[3]Kay Gimm is a group leader at the Institute of Transportation Systems in DLR (German Aerospace Center), Braunschweig, Germany Kay.Gimm@dlr.de

TABLE I
CHARACTERISTICS OF BIG DATA

	Characteristic	Description	Does it apply to the case?
1	Volume	Quantity of generated and stored data	• Trajectory data: 3.5 GB per 12 daytime hours. • Generated interaction data: 1.4 million pairs per 12 daytime hours. • Video data: 15.5GB per hour
2	Variety	Varied type and nature of the data.	Not strictly the case because the data is a conventional Relational Database with various tables.
3	Velocity	Speed at which the data is generated and needs to be processed.	Need to process, extract and model many different scenarios in real time.
4	Veracity	Truthfulness or reliability of the data.	The data from the AIM Research Intersection is accurate enough for the purposes of traffic research.
5	Value	Worth in analyzing large datasets.	• Intrinsic benefit in having more data for research. • Rare events will be found.

encountering a crossing bicycle. A scene would be a specific red car turning right at a specific intersection and crossing with a specific cyclist in a certain way.

The presented techniques and methods will be applied with a scenario-based approach and following the KDD method to develop the collaborative framework that solves the Big Data problem of automatically and collaboratively mining several different traffic scenarios from a sensorized infrastructure.

III. METHODOLOGY

In this section, the collaborative framework developed by the German Aerospace Center (DLR) for a semi-automatic and efficient the extraction and analysis of relevant traffic situations is described. Semi-automatic in this context refers to the fact that the user must manually define the conditions for criticality and atypicality for each scenario.

A. Design of the framework

The essence of the framework is to define in a standard way the scenarios to be mined , extract the candidate scenes as well as the relevant properties of each trajectory and using those data, filter the scenes to get the ones that match the target scenario. Each generated intermediate result is saved to a database allowing to reuse them for other scenarios and users (collaboration). Special kind of scenarios that part from already mined scenarios can be used for increased efficiency.

The architecture of the proposed framework is shown in "Fig. 1". The input data of the pipeline is the source dataset (node 1 in orange) which contains trajectory data (time series data with one timestamp and object ID for each row). The output data are the scenes (nodes 4.x in yellow), models (node 6 in purple) as well as the critical (node 7 in red) and the atypical (node 8 in green) scenes of the mined scenarios.

From the dataset, the *TrajectoryDataMiner* generates the trajectory data for latter steps:

- Time range: start and end time for each trajectory.
- Object class: class of the object (pedestrian, car, bike, van, truck, etc.).
- Route: sequence of road lanes of the digital map (e. g. the AIM Research Intersection "Fig. 2") that the trajectory took in its course. This is calculated using the map matching algorithm presented in [9]. Because this is computationally expensive, the trajectories are down sampled beforehand.

Using the time ranges data from the trajectories, the *SceneMiner* generates the existing scenes. In this context, a scene is a combination of a given number of trajectories that coexist in time.

Finally, the *ScenarioMiner* filters the scenes using the *TrajectoryData* to get the ones that belong to the scenario. A scenario is an abstract traffic situation that is composed of trajectories interacting in a certain way. To filter the scenes, the following properties of the scenario may have been included in its definition:

- Object class for each trajectory.
- Route that each trajectory follows.
- Interactions between pairs of trajectories according to [10].
- Region of interest (ROI) where all the trajectories and their interactions exist.
- Special conditions that a scene has to fulfill to belong to the scenario. For example, a PET (Post-Encroachment Time) smaller than 5 seconds between trajectories 1 and 2.

There are other types of scenarios (*SubScenario*, *MultipleScenario*, *CombinedScenario*) whose scenes are mined from the scenes of their related scenarios to be more efficient.

A *SubScenario* is a scenario that is contained in another parent scenario. The *SubScenario* is more specific and restrictive than the parent scenario. All the scenes that belong to the *SubScenario* must also belong to the parent scenario. The pipeline to mine the scenes for a *SubScenario* consists of a single *SubScenarioMiner* that filters the scenes from the parent scenario to get the ones for the *SubScenario*. It works exactly the same as the *ScenarioMiner* previously defined.

A *MultipleScenario* is a scenario that is a multiple simultaneous instance of a parent scenario. Optionally, one or more of the trajectories may belong to all the converging scenes (they are the intersection between the multiple scenes). The *MultipleScenario*'s scenes must have more trajectories than the parent scenario. All the scenes that belong to the *MultipleScenario* are composed of intersections of scenes of its parent scenario. All the trajectories in the *MultipleScenario* must coexist in time with all others. To mine the scenes of a *MultipleScenario* only a *MultipleScenarioMiner* is needed

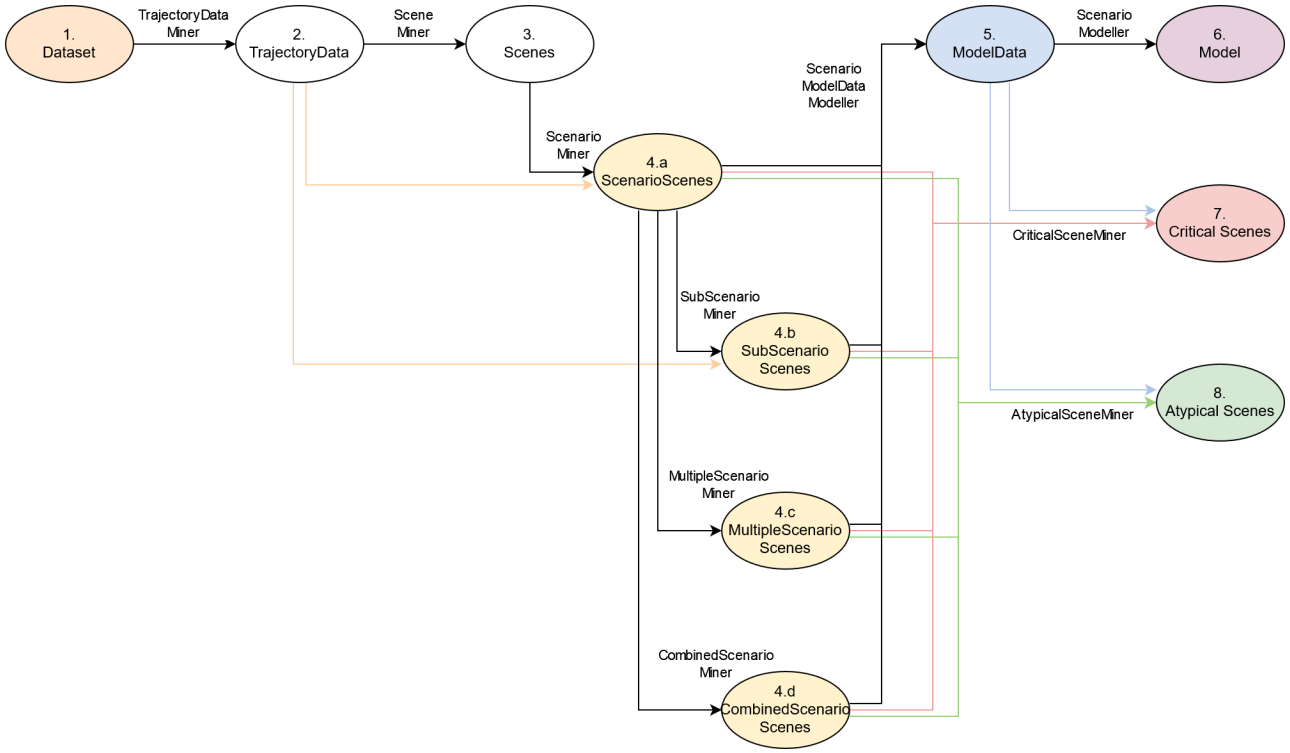

Fig. 1. Scenario mining & modelling pipeline.

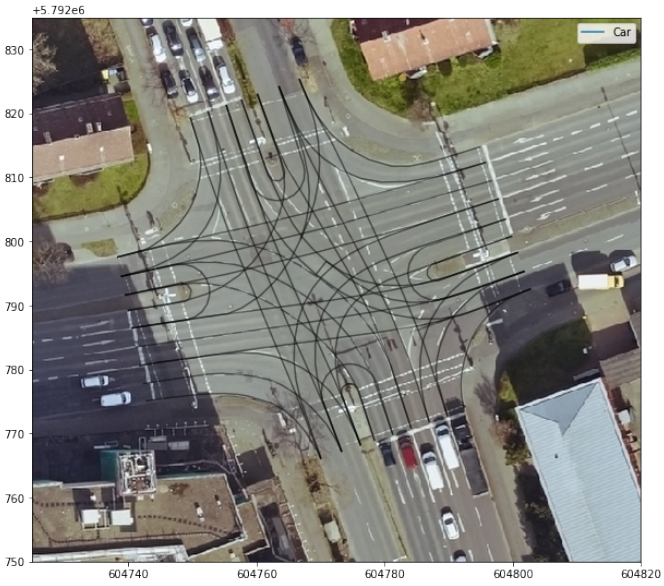

Fig. 2. Digital map on satellite image of the AIM Research Intersection, Braunschweig (Germany).

which combines the coexisting scenes from the parent scenario to get the repetitions for the *MultipleScenario*.

A *CombinedScenario* is a scenario that is a combination of two or more parent scenarios that happen simultaneously. Some trajectories may belong to more than one of the scenarios. The *CombinedScenario* must have a smaller or equal number of trajectories than the sum of the parent scenarios. All the scenes that belong to the *CombinedScenario* are composed of intersections of scenes from its parent scenarios. The pipeline to mine the scenes of a *CombinedScenario* consists of a *CombinedScenarioMiner* that combines the coexisting scenes from the parent scenarios to get the combinations for the *CombinedScenario*.

After mining a scenario (*Scenario*, *SubScenario*, *MultipleScenario* or *CombinedScenario*), the pipeline to model it consists of a *ScenarioModelDataModeller* and a *ScenarioModeller* as can be seen in "Fig. 1". The first one generates the model data of the scenario from its scenes. The model data consists of the values of the modelling variables defined for the scenario and aggregated by their dimensions. For example:

- By [time, object ID]: velocity magnitude, longitudinal acceleration, lateral distance to lane, etc.
- By [time, object ID 1, object ID 2]: TTC (Time To Collision), distance, relative velocity, ePET (expected PET), etc.
- By [object ID]: maximum velocity magnitude, minimum longitudinal acceleration, etc.
- By [object ID 1, object ID 2]: PET (Post-Encroachment Time), minimum TTC, minimum ePET, etc.

The *ScenarioModeller* creates valuable models from the scenario's model data such as histograms, correlations and other metrics of the variables, as defined by the user for the

scenario.

After having mined and modelled a scenario, its critical and atypical scenes may be mined.

To mine the critical scenes from a scenario a *CriticalSceneMiner* is needed. The miner performs a filtering of the scenario's scenes based on the scenario's *ModelData*. The conditions are defined by the user but the default condition for criticality is either a PET absolute value smaller than 1 second [1] or a TTC smaller than 1 second with a distance smaller than 10 meters.

To mine the atypical scenes from a scenario an *AtypicalSceneMiner* is needed. The miner performs a filtering of the scenario's scenes based on the scenario's *ModelData*. The exact conditions which a scene must fulfill to be considered atypical should be defined by the user. Criteria such as the Hausdorff distance from the typical historical trajectory are encouraged but a simple default condition is provided which can detect atypical situations. The default condition for atypicality is that at least one of the model variables of the scene is atypical. A variable is atypical when at least 20% of its values are in the top or bottom 0.5% of the historic distribution.

B. Features of the collaborative framework

This framework has certain features that are of special interest for our application:

- Mining principles:
  - "Lazy" filtering: start with the faster calculations that reduce the amount of data the most to avoid latter filtering calculations.
- Distributed services approach:
  - Each *Miner/Modeller* is responsible for a service.
  - Each *Miner/Modeller* loads the data it needs from the database and saves its results to its corresponding table in the database.
  - Each *Miner/Modeller* checks whether its target data was already mined before generating it.
- Collaboration:
  - A single application mines all scenarios for all researchers.
  - Reuse calculations across researchers and scenarios.

The presented framework allows to automatically mine and model traffic scenarios in a standard, scalable and collaborative way, in contrast to the approaches seen in the literature. "Table II" describes the problem presented in the introduction and how the framework solves them.

TABLE II

PROBLEMS SOLVED BY THE FRAMEWORK

	Problem	Solution
1	Large amounts of data to analyze at real-time speed.	1) Big Data techniques. 2) Efficient data mining techniques. 3) Scalable software.
2	Duplicated calculations across scenarios and users of the pipeline.	Collaborative mining: central database where the mining calculations are shared.
3	Varied scenarios requiring different mining algorithms.	Single application that is valid for all users for all needed scenarios.

IV. RESULTS

An application was developed based on the described framework and was validated experimentally. A team of researchers studied eight scenarios (see "Table III") mined from four days of data. Figure "Fig. 3" shows the network of scenarios, their scene counts and how they relate to each other. The arrows go from parent scenarios to children scenarios.

TABLE III

DESCRIPTION OF MINED SCENARIOS

index	***scenario_name***	***description***
S1	bike_MRU_crossing	Bike crossing with an MRU with a PET smaller than 5 seconds.
S2	north_bike_RT	North crossing bike crossing with a right turning MRU.
S3	north_bike_LT	North crossing bike crossing with a left turning MRU.
S4	north_bike_RT2	North crossing bike crossing with two right turning MRUs.
S5	bike_truck_crossing	Bike crossing with a truck with a PET smaller than 5 seconds.
S6	north_bike_RT_truck	North crossing bike crossing with a right turning truck.
S7	north_bike_LT_truck	North crossing bike crossing with a left turning truck.
S8	north_bike_RT_comb_LT	North crossing bike crossing with a right and a left turning MRU.

5,569,949 scenes of two trajectories are found in the dataset, from which 5,070 belong to scenario 1 (S1) consisting of a cyclist and an MRU (Motorized Road User) that cross with a PET smaller than 5 seconds, which is a data reduction ratio of 1100:1. This first scenario's function is to reduce the volume of the data before performing more expensive filtering calculations.

From these 5,070 scenes, 822 are cyclists crossing the north street of the intersection that cross with a right turning MRU (S2). Another 404 are the same case but with a left turning MRU (S3).

From these scenarios (S2 and S3), 89 situations are found where the north street crossing cyclist crosses with a right turner and a left turner simultaneously (S8). From S2, 77 scenes are found where the north crossing cyclist crosses with two right turners simultaneously (S4).

Further 212 scenes are north crossing cyclists that cross with a truck (S5). 16 are right turning trucks (S6) and 14 left turning trucks (S7).

Experimental results using a Intel(R) Xeon(R) Gold 6126 CPU @ 2.60GHz processor with 12 cores and 64 GB of RAM proved that mining 12 hours of real data only takes 3 hours and 9 minutes. Almost all the run time is spent in generating the *TrajectoryData*, candidate *SceneData* and getting the first scenario's *SceneData*. The rest of the

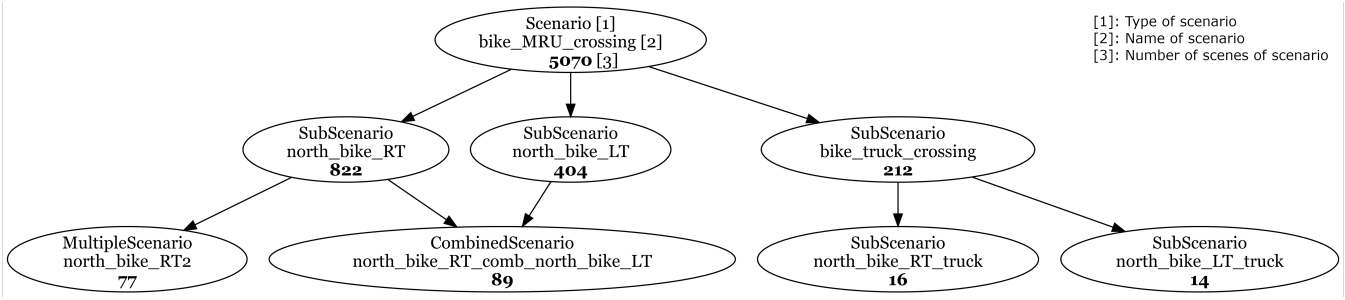

Fig. 3. Network of mined scenarios.

scenarios part from the small amount of scenes found for the first one and their mining time is negligible in comparison. This means that the run time is independent of the number of users and scenarios to be mined for all practical purposes.

V. Discussion

The accuracy of the first scenario's mining is dependent on two factors. First, the reliability of the object classification algorithm to detect bicyclists. Second, the reliability of the PET not to filter out interesting situations, which is why the selected PET threshold (5 seconds) is rather generous.

The accuracy of the mining of the next scenarios is dependent on the the capability of the object classification algorithm to detect cars and trucks, as well as the reliability of the map matching algorithm developed for the framework presented in this paper.

To further study the accuracy of the pipeline, the mining of the scenes for the *SubScenario north bike vs RT* (S2) from the scenes of the scenario *bike vs MRU* (S1) was validated for half a day of data (458 scenes) by reviewing the scenes' videos.

The confusion Matrix for the *SubScenario north bike vs RT* for a day's data is shown in "Table IV". In this confusion matrix, *True* means that a scene from S1 was correctly identified as belonging to S2, and *False*, the contrary. From the 100% of the candidate scenes 19.1+0.9=20% are classified as belonging to the target scenario. 19.1% of the scenes are correctly identified (True Positives: green) but 0.9% are falsely classified as belonging to it (False Positives: red) and no scenes are falsely classified as not belonging to the scenario (False Negatives: orange). The remaining 80.0% are True Negatives in yellow.

TABLE IV

Confussion Matrix of *ScenarioMiner* for scenario *north_bike_RT* and time range 2022-10-19 6:00 - 13:30.

Confussion Matrix	labelled True	labelled False
is True	19.1%	0.0%
is False	0.9%	80.0%

An example of a true positive situation can be seen on the satellite map in "Fig. 4" and as a video in "Fig. 5".

The false positives are caused by pedestrians or bicycles that are recognized as MRUs and are falsely classified as

Fig. 4. Example of True Positive of *north_bike_RT* (S2).

Fig. 5. Example of True Positive of *north_bike_RT* (S2).

right turners. The error in map matching is due to the fact that it considers the closest road lane as the path the trajectory took and because the walkway is not a part of the digital map, an object performing a right turn on the walkway will be classified as a right tuner. See an example in "Fig. 6", where the bicycle with a front cart was falsely classified as a right turning car. This can be solved by addressing the object recognition software errors and by adding lanes for the walkways in the digital map of the intersection.

Some of the true negatives occur due to erroneous tracking of the objects or atypical trajectories that don't match the specifications of the scenario in question. See examples in "Fig. 7" and "Fig. 8". In the first one, only the end of the trajectory of the right turner was recognized. The second one shows an atypical trajectory from a bicycle which deviates from the designated route. If these cases should actually be

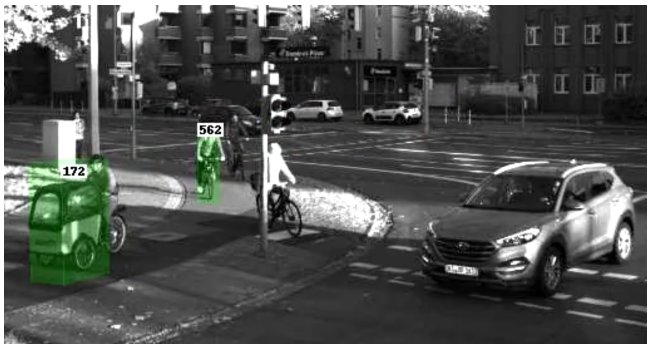

Fig. 6. Example of False Positive of *north_bike_RT* (S2).

detected as valid scenes, the scenario should be adjusted accordingly.

Fig. 7. Example 1 of possible False Negatives of *north_bike_RT* (S2).

The accuracy problems that have been detected in the scenario mining pipeline are not critical for the objective of generating valuable traffic scenario data from a big data source and are unrelated to the presented collaborative framework itself. The number of false positives is actually very limited. No false negatives were detected for the validated data.

VI. Conclusion

This paper presents a framework for the semi-automatic mining of traffic scenarios in a standard, scalable and collaborative way. It was implemented as central application and used by a team of researchers to mine and analyze 8 scenarios. This proved that the framework was an efficient tool for data mining and basic modelling which allows the researchers to focus on the final data analysis.

Nevertheless, this application should be further developed and tested for more scenarios as well as more SMoSs be implemented for the analyst to use. Universal methods to detect critical and atypical scenes are required for a fully automatic scenario mining framework. The research by Oksana Yastremska-Kravchenko, Aliaksei Laureshyn, et al. [12] linking human-perceived severity to objective and quantitative variables could be an appropriate starting point for this matter. If these universal objective variables could be generalized for all scenarios it would be sufficient to calculate them and define an appropriate threshold to automatically find the critical situations. A way to define the scenarios based on maneuvers would be of special interest, as that is a standard of the automotive industry [13], especially for simulation.

Fig. 8. Example 2 of possible False Negatives of *north_bike_RT* (S2).

Acknowledgment

I would like to thank my DLR colleagues Marek Junghans, Peter Wagner, Juan Trullos and especially Louis Calvin Touko Tcheumadjeu for their invaluable assistance in the writing of this paper in accordance with the highest academic standards.

References

[1] Knake-Langhorst, S., & Gimm, K., AIM Research Intersection: Instrument for traffic detection and behavior assessment for a complex urban intersection. Journal of large-scale research facilities JLSRF, 2, A65-A65, 2016

[2] Piotr Olszewski, Aliaksei Laureshyn, InDev Review of current study methods for VRU safety.

[3] K.P. Petr Pokorny, Observations of truck-bicycle encounters: A case study of conflicts and behaviour in Trondheim, Norway, 2018.

[4] G. Andrienko, N.Andrienko,P.Bak,D.Keim,S., A conceptual framework and taxonomy of techniques for analyzing movement, 2011.

[5] Yu Zheng, Trajectory Data Mining: An Overview, 2015.

[6] W.F. Gregory Piatetsky-Shapiro, Knowledge Discovery in Databases, 1991.

[7] George Firican, The 10 Vs of Big Data. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx.

[8] John A. Volpe, Pre-Crash Scenario Typology for Crash Avoidance Research, 2007.

[9] Lars Klitzke, Clemens Schicktanz, et al., Scenario Mining in the Urban Domain: Exploiting the Topology Scenario Mining in the Urban Domain: Exploiting the Topology of a Road Network for Maneuver Annotation and Scenario Extraction, 2022.

[10] G. Markkula, R. Madigan, et al., Defining interactions: a conceptual framework for understanding interactive behaviour in human and automated road traffic, 2020.

[11] Varhelyi, A, Drivers' speed behaviour at a zebra crossing: a case study. Accident Analysis & Prevention, 30(6), 731-743, 1998

[12] Oksana Yastremska-Kravchenko, Aliaksei Laureshyn, et al., What constitutes traffic event severity in terms of human danger perception?, 2022

[13] ASAM, OPEN Scenario. https://www.asam.net/standards/detail/openscenario/.