https://doi.org/10.2352/EI.2023.35.16.AVM-119

Design of an Automotive Platform for Computer Vision Research

Dominik Schörkhuber[1]***, Roman Popp***[2]***, Oleksandr Chistov***[3]***, Fabian Windbacher***[4]***, Michael Hödlmoser***[4]***, Margrit Gelautz***[1]***;*** [1] ***TU Wien,*** [2] ***ZKW Lichtsysteme GmbH,*** [3] ***ZKW Group GmbH,*** [4] ***emotion3D GmbH; Vienna, Austria***

Abstract

The goal of our work is to design an automotive platform for AD/ADAS data acquisition in view of subsequent application to behaviour analysis of vulnerable road users. We present a novel data capture platform mounted on a Mercedes GLC vehicle. The car is equipped with an array of sensors and recording hardware including multiple RGB cameras, Lidar, GPS and IMU. For future research on human behaviour analysis in traffic scenes, we compile two kinds of data recordings. Firstly, we design a range of artificial test cases which we then record on a safety regulated proving ground with stunt persons to capture rare events in traffic scenes in a predictable and structured way. Secondly, we record data on public streets of Vienna, Austria, showing unconstrained pedestrian behaviour in an urban setting, while also considering European General Data Protection Regulation (GDPR) requirements. We describe the overall framework including the planning phase, data acquisition and ground truth annotation.

Introduction

Computer vision research in the automotive industry typically involves the design and development of systems and algorithms that enable vehicles to perceive and understand their surroundings. This can include tasks such as object detection and classification, scene understanding, and motion planning. One approach to designing an automotive platform for computer vision research is to use a vehicle equipped with sensors and cameras that can provide a rich and diverse dataset for training and evaluating machine learning algorithms. This may include Lidar, Radar, and/or camera sensors, as well as on-board computing resources for running algorithms in real-time.

We target our research towards systems that have the potential to improve the safety of vulnerable road users. In a report by the NHTSA (National Highway Traffic Safety Administration, USA) [1], traffic accidents are recorded and analyzed. While the number of injuries and fatalities decreases for most years, since the year 2000, the proportion of fatalities outside the vehicle[1] has steadily increased from 20% in 2000 to 34% in 2020. This indicates that more emphasis needs to be put on the protection of non-occupants in traffic scenarios. While the outcome of an accident can be fatal, their statistical occurrence is minor. According to [2], vehicles typically run several billion kilometers per accident, making data collection of rare events largely ineffective, or even unfeasible. However, for the development of data-driven machine learning models, as they are widely used in automotive applications, there is a requirement for vast amounts of data for training and testing such models. While data from public driving datasets is abundantly available, the application of such datasets is often hindered by the specifics of the dataset contents such as the cameras used and situations recorded, increasing the domain gap between the field of application and data for training and validation during development. To tightly control the parameters of a dataset we therefore opt to create our own platform for data acquisition. A key aspect of the design is to have a robust architecture that can support the development and testing of various computer vision algorithms. Overall, the design of an automotive platform for computer vision research should aim to provide a flexible and scalable solution that can support a wide range of research and development activities in the field of autonomous vehicles. We summarize the contributions of our work as follows:

- We present a holistic approach to a vision-based automotive platform from sensor setup to data recording.
- We describe the hardware design and decisions for trade-offs between flexibility and stability of the system.
- We tailor our requirements mainly towards computer vision research, while also considering restrictions imposed by privacy and data protection laws.

In the remainder of the paper, we discuss the integral parts of the developed platform for automotive computer vision research. We start by introducing our *sensor platform*, consisting of a mounting system, imaging sensor and electronics. We discuss their integration on the hardware level and their purpose. We continue to elaborate on how we planned *driving scenarios* for recording to not only record data in the wild but also be able to construct repeatable test cases with defined interactions between pedestrian and driver. Especially on public streets, data recording is restricted by law. We therefore discuss the *General Data Protection Regulation (GDPR)* and how it affects our work. Lastly, we describe the tools and workflow used to label the recorded data for behaviour analysis of vulnerable road users.

Related Work

Recording and curation of data is a vital aspect of computer vision research. Data is collected and published in many work concerned with research in the field of autonomous vehicles, but the platforms used to record data and the recording process itself are often not addressed. In this work, we aim to improve on that and present our approach to data recording and data curation for research and observation of pedestrian behaviour from an ego-vehicle. Only few datasets tailored to pedestrian behaviour analysis are currently publicly available, e.g. [3, 4, 5, 6]. While transfer learning is a common technique in deep learning and computer vision, [7] has shown that it has only limited application for pedestrian behaviour analysis. When obtaining real data is difficult, synthetically generated data [8] can be an alternative. While realistic simulation of visual appearance for actors in automotive settings is possible [9], synthesizing realistic human behaviour is extremely challenging.

[1] motorcyclists, pedestrians, bicyclists, and other non-occupants

As a vital component of computer vision systems, the **sensor setup and used hardware components** are only sometimes described in the literature. In [10], the authors briefly review the used sensors, their respective purpose and placement. While most previous work puts its emphasis on the recorded data and algorithms (e.g., [11]), components beyond sensors such as recording hardware are typically not discussed. Contrary to a real-time capable system like the one presented in [12], our work puts its focus on developing a recording platform.

The **description of scenarios** is an important part of automotive development and testing. Depending on the recorded situation and usage of sensor data, differently detailed scenario descriptions are required. In [13], the authors describe scenario-based safety assessment for simulation purposes in a formal framework. On a wider scope, in [14] the terms *scene*, *situation* and *scenario* for automotive testing are substantiated. Under the definition given in this work, we plan scenarios which are defined by actions and events to reach a defined goal.

In the context of **image and video annotation**, computer vision labeling is the process of adding metadata to an image or video in order to provide additional information about its content. This is often done using specialized software tools that use artificial intelligence and machine learning algorithms to automatically analyze the content of an image or video and generate labels that describe its key features. For the annotation of images, a range of tools exists, e.g., [15, 16, 17]). A comprehensive list of tools is presented in [18, 19]. However, while video frames could be annotated as separate frames, the incorporation of temporal dependency in the annotation tool enables more efficient annotation. For the purpose of behaviour analysis and object tracking, we aim to generate spatio-temporal annotations, i.e., bounding boxes linked between frames. Only few free and open source tools are available for this [20, 21, 22, 23] and even fewer allow for semi-automatic labeling by single object tracking [24, 19] which is required to annotate objects over time efficiently.

Method

In this section, we discuss our steps to building the automotive platform for data recording. We go into detail on design decisions, hardware choices and the electrical design of components. We continue to elaborate on how and where the recorded data is stored, and what precautions are necessary to store data safely. We pay particular attention to capture scenarios of interest and to minimize the waste of resources. Furthermore, we present our setup of different scenarios that were recorded on public streets and on test tracks. Lastly, we go into detail about the annotation of video data, and present our approach to annotating vulnerable road users in videos efficiently.

Sensor Platform

For data acquisition, an off-the-shelf car was equipped with additional hardware. The car model used is a Mercedes GLC, with an additional mounting rail that was constructed and attached to the front of the vehicle as shown in Figure 1. While other systems (e.g., [11, 10]) mount their sensors on top of (and around) the vehicle, this particular design was chosen to mimic the positioning of headlight systems and to allow for easy mechanical access to sensors during prototyping. The intention is to have a prototypical setup for a later integration of sensor components inside the headlamps. For reasons of traffic safety, the rail was equipped with additional turn signals, the number plate and reflectors.

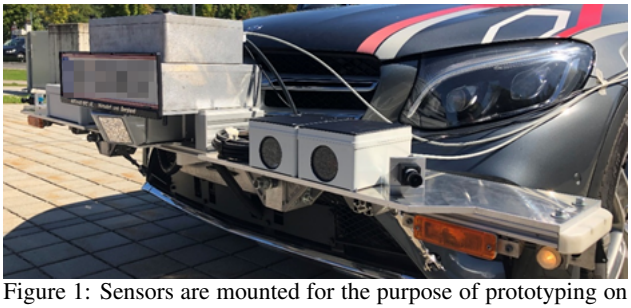

Figure 1: Sensors are mounted for the purpose of prototyping on an easily accessible rail in front of the car.

In total, we use five imaging sensors oriented towards the front and sides of the vehicle as visualized in Figure 2. The side-facing sensors allow for a 180° vision cone in front of the vehicle. Especially vulnerable road users approaching from the sides are recognizable early with these sensors. Further we have mounted a camera with a Red Clear Clear Blue (RCCB) filter for low-light scenarios, a high definition camera with narrow field of view to detect objects at range and and a thermal camera oriented towards the front. Besides these cameras, we use a Lidar sensor for 3d perception. In addition to the imaging sensors, an inertial measurement system is mounted on the vehicle.

The used sensors require different measures of protection. While the side-facing cameras and the Lidar are water proof, the front-facing RGB, RCCB and thermal cameras need special housings for protection from water. In principle, the RCCB and high definition RGB camera could also be placed behind the windshield of the car, the housing of the thermal camera has the additional requirement of not blocking any thermal waves.

For the electrical connection of different sensors, different interfaces are used. An overview of the sensors, components and their connections is shown in Figure 3. A widely used interface in the automotive industry is Gigabit Multimedia Serial Link version 2 (GMSL2). To feed the GMSL2 sensor signal to the recording PC, we first deserialize the sensor signal and convert to USB or Ethernet. Depending on the sensor component, we either use the B-PLUS MDILink (Measurement Data Interface)[2] or GMSL2-to-USB deserializers. The used sensor components and some of their properties are listed in Figure 2. As Lidar we use the Luminar[3] H3 Lidar, which supports Ethernet connectivity natively. Likewise, the front-facing high definition RGB camera (Basler[4] ac2040-35gc) supports an Ethernet stream for data. The two side-facing cameras are Leopard Imaging LI-IMX490 sensors, which provide a GSML2 interface and connect to the MDILink. The other sensors like the thermal camera (Infiray Xsafe II M6S) and the RCCB camera (Leopard Imaging LI-IMX424-GMSL2) are connected via GMSL2-to-USB deserializers.

All these sensor inputs are combined on an embedded PC in the vehicle's rear trunk. A specialized software framework, provided by Lake Fusion Technologies GmbH[5], is responsible for

[2] https://www.b-plus.com/en/products/automotive/vehicle-data-harvesting/data-interfaces-and-converters/mdilink-gmsl2

[3] https://www.luminartech.com/

[4] https://www.baslerweb.com/

[5] https://lf-t.net/

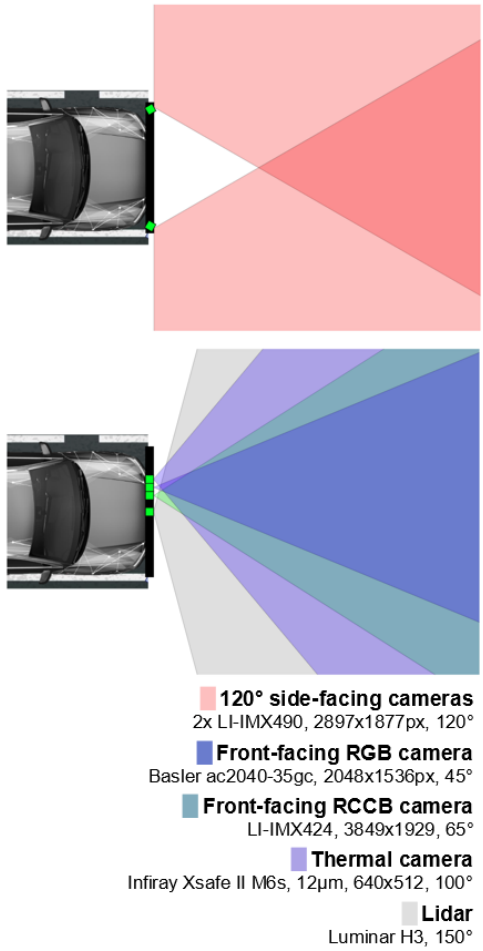

Figure 2: Visualization of the field of views of used imaging sensors. The side-facing cameras (top image) form a 180° detection area in front and to the sides of the vehicle. The other sensors (bottom image) are oriented towards the front.

data serialization to the disks. The system further integrates an inertial measurement unit (IMU) mounted on the test vehicle. The IMU consists of a three axis gyroscope, a three axis accelerometer, a three axis magnetometer and a temperature sensor as well as a global navigation satellite system receiver. Synchronization of the sensor signals is done via timestamps. Some of the sensors integrate their own timeserver and in these cases the sensor itself adds a timestamp, for other sensors, the recording software adds a timestamp. For synchronization of the different time servers the precision time protocol (PTP) is used and the GPS time of the IMU is set as reference. The recording software provides a web based user interface (UI) using OpenHAB[6]. The UI is shown in Figure 4. It displays the current status of the system and allows starting/stopping of the recording as well as tagging of the data during recording. Additionally it provides the functionality for configuration of the sensors.

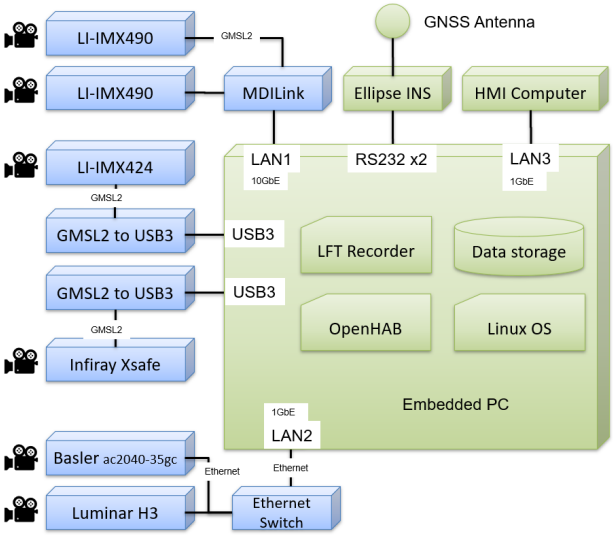

Figure 3: The recording system includes a range of sensors connected to an embedded PC to synchronize and record the sensor signals. Furthermore, a HMI (human machine interface) computer allows control and supervision of the system during operation.

Data Privacy and Protection

Effective since 2018, the GDPR[7] has set a precedent for data protection in the European Union. Data recording for autonomous driving research inevitably captures personal data of people surrounding the vehicle such as the faces of pedestrians and cyclists or number plates of cars. While imaging sensors like (RGB) cameras and Lidar capture human traits and actions in an identifiable way, thermal or depth cameras for example can provide a privacy-preserving alternative in some cases. However, for the application of behaviour understanding of vulnerable road users, an image representation of the non-verbal communication between vulnerable road user and driver is usually necessary. Defined by the GDPR, the data subject has a number of rights. The recording of personal data must be communicated to the data subjects in a clear and understandable way. For this purpose, our recording vehicle is equipped with a large sign showing contact information and a QR-Code leading to a statement of data policy. The data subject has further a right of information and access to the data. If a person was recorded, he/she can, at any point, request the deletion of their data. It is the data controller's obligation to process and respond to such requests. To ensure confidentiality and enable processing of data requests, we document any access to the recorded data. One possible way to ensure the protection of personal data is anonymization. Proper anonymization of any identifiable trait would effectively render the data outside of the scope of the GDPR. Since this is too labor intensive and because anonymization would hinder the analysis of non-verbal communication between humans, we refrain from anonymization and fall back to pseudonymization for data protection, which ensures that any piece of data cannot be attributed to an individual anymore without additional information. By this definition, we achieve pseudonymization by encryption of the data.

[6] `https://www.openhab.org/`

[7] `https://gdpr.eu/`

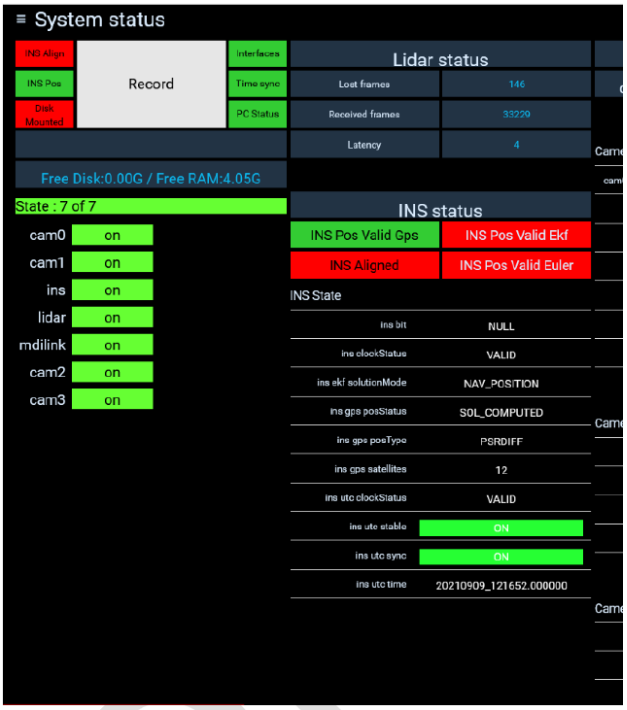

Figure 4: The human computer interface allows for control and supervision of the recording system. Free disk space, mounted disks, free RAM and the INS status as well as successful connection to the sensors are supervised during recording.

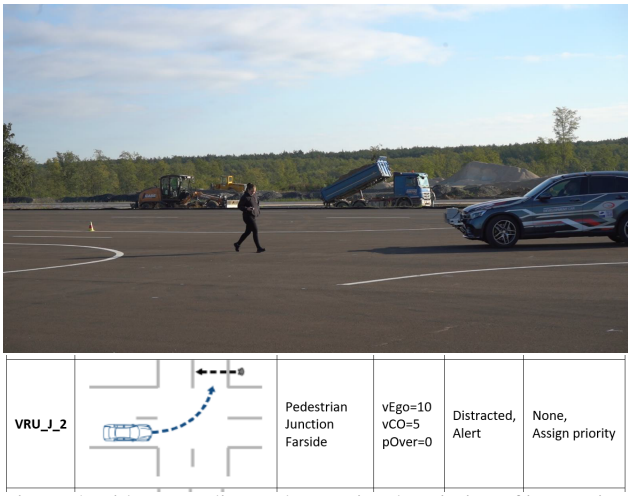

Figure 5: Video recording and scenarios description of interaction between pedestrian and vehicle on an automotive proving ground.

Scenarios

Vulnerable road users, such as pedestrians, cyclists, motorcyclists, and people with disabilities are individuals who are more likely to be involved in a traffic accident due to their physical vulnerability. These individuals are at a higher risk of injury or death in a collision compared to other road users, such as motorists who are protected by the structure of their vehicle. While it is important for all road users to be aware of and take precautions to protect vulnerable road users, appropriate measures also need to be integrated in the design of assisted and autonomous driving platforms. Non-verbal communication between a driver and pedestrian can be important for maintaining safety and order on the road [25, 26]. There are several ways that a driver and pedestrian can communicate non-verbally. Making eye contact with a pedestrian can indicate that the driver has seen them and is aware of their presence. This can be particularly important when pedestrians are crossing the road. Drivers can use hand signals to indicate their intentions to pedestrians. For example, a driver might wave their hand to indicate that it is safe for a pedestrian to cross the road. Drivers can use their headlights, turn signals, and other vehicle signals to communicate with pedestrians. For example, flashing headlights might be used to indicate to a pedestrian that it is safe to cross the road. Drivers can also communicate non-verbally through facial expressions. To complicate things even more, the meaning of non-verbal signals may vary with geographic location or cultural background.

To record such interactions in large number, we follow two strategies. Firstly, we design artificial test cases on an automotive proving ground. With the foreseen scenarios, we aim to verify methods for object detection, object tracking and behaviour prediction. Our scenarios are based on the insight that when the path of vehicle and pedestrian are going to cross, communication must be established and the right of way is negotiated. The scenarios take place at a four-way crossing, as depicted in Figure 5, or straight parts of the track. In our planning, the test cases are sketched following a description of the vehicle's and pedestrian's path, while several variables describe their behaviour. Most importantly, we pre-define for each test case whether the pedestrian is paying attention to the scene or if they are in a distracted state. The latter may be represented by the pedestrian either looking away from the vehicle or being distracted by a smartphone. Another variable describes who has the right of way, defining if either the pedestrian or the vehicle has to stop.

To implement custom traffic scenarios in a safe manner, two stunt people were hired. One acted as a pedestrian with the intention of crossing the road, and a second professional drove the car. With their experience, the stunt persons were able to execute the defined scenarios with utmost precision and their previous experience in the entertainment industry was helpful to present realistic forms of non-verbal communication. The driver had to establish and maintain a defined speed and be able to stop the vehicle at a specified point. Additionally, the pedestrian also needed to adhere to the right starting point and timing, such that both meet at the desired point. For easier orientation, some tests keypoints on the test track were marked with traffic cones, and communication between a coordinator and both stunt people was maintained via radio phones.

Secondly, we recorded real driving sequences by driving through urban areas. With the goal of meeting large numbers of pedestrians, we aimed for the inner city of Vienna, Austria. The most interesting sequences we obtained are situations where the right of way must be negotiated, since it is more challenging to recognize the intention of a pedestrian wanting to cross the street when there is no indication of a safe way of crossing, i.e. a cross-walk or traffic lights. With this in mind, we aimed mostly for narrow streets, where said situations are more likely to arise. While conducting the recordings, tasks were distributed among four team members inside the vehicle. The driver was solely re-

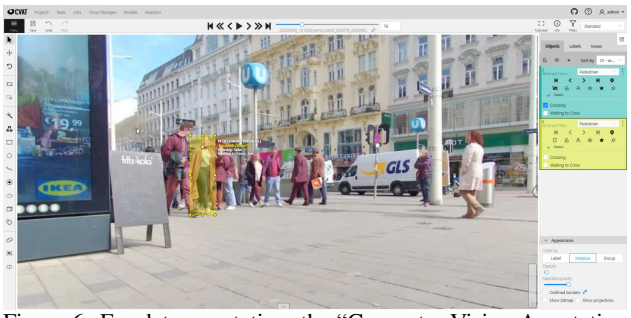

Figure 6: For data annotation, the "Computer Vision Annotation Toolkit" (CVAT) was used. To facilitate fast and efficient annotation, we integrated a state-of-the-art single object tracking algorithm for semi-automatic annotation.

sponsible for steering the car, while the co-driver navigated to streets matching our desired traffic profile. Two more people were in charge of the recordings controlled by the OpenHAB interface; one had to start and stop recordings while the other was tagging the occurrences of vulnerable road users.

Video Annotation

The Computer Vision Annotation Toolkit (CVAT) is a free and open-source software that enables users to annotate images and videos for computer vision tasks such as object detection, segmentation, and tracking. CVAT offers a range of features that make the annotation process efficient and accurate, including the option to review existing annotations and make use of interactive annotation modes, such as support for semi-automatic annotation. In CVAT's semi-automatic video annotation workflow, an algorithm is used to automatically identify and label specific objects or scenes in a video. This is typically done via machine learning algorithms and pre-defined rules or criteria. Once the algorithm has identified and labeled the objects or scenes in the video, a human annotator can review the labels and make any necessary adjustments or corrections. This allows for a more efficient and accurate video annotation process, as the computer is able to handle the initial identification and labeling, while the human annotator can focus on ensuring the accuracy and quality of the annotations. For semi-automatic annotation, CVAT supports the integration of interactors, detectors and single object trackers, where we are focusing on the latter. Single object tracking mechanisms are able to track image templates from frame to frame to ultimately form the trajectory of an object. The architecture of CVAT separates the integration of 'AI Tools' into a containerized environment managed by Nuclio, where the required algorithms are provided as serverless functions which communicate with the CVAT application through a web-based interface. We integrate the state-of-the-art tracking algorithm TransT [27], which is class agnostic and sufficiently accurate, into the annotation system's architecture. The integration of TransT into CVAT has been made publicly available. We show an exemplified usage of CVAT in Figure 6.

Conclusions

In this paper, we discussed our approach to an automotive platform for computer vision research. We elaborated on the sensors and hardware components used and presented our methods

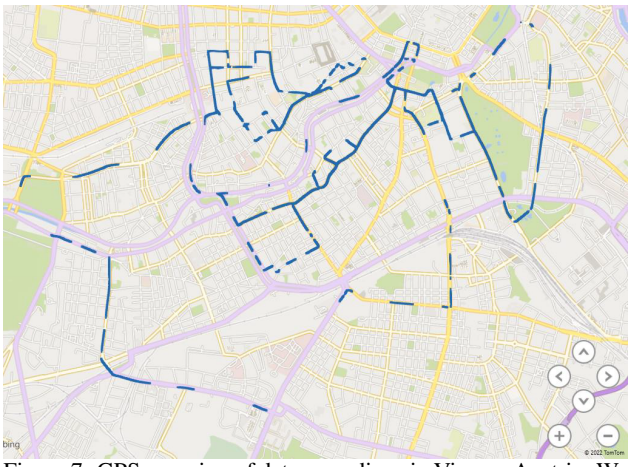

Figure 7: GPS mapping of data recordings in Vienna, Austria. We enable data recording during our drives for specific time periods corresponding to areas where we expect a significant number of vulnerable road users.

to record data for pedestrian behaviour analysis. We found that while the sensor placement on the rail allowed for easy mechanical access, the construction has some structural issues, as vibrations are more strongly translated to the sensors. Another aspect we evaluated while recording on public streets was the system's resilience. While mechanical perturbation of the vehicle had no impact on the sensor recordings, we found that the side-facing cameras occasionally showed dropped frames, which we attribute to bandwidth limitations. An issue observed was the signal alignment of the INS, which worked as expected when the vehicle was moving but abnormal vehicle movement like abrupt breaking affected the INS negatively. During recording we therefore supervised the INS status as shown in Figure 4. With respect to regular driving maneuvers, interaction with pedestrians is comparably rare and requires a data acquisition strategy to capture scenarios of interest efficiently. To balance between realistic behaviour and data quantity, we chose to record on public streets and automotive proving grounds. Recording on public streets is arguably less resource intensive than the usage of proving grounds, but also requires a considerable amount of personnel to drive, navigate, supervise, select and tag data as described in the previous section. Figure 7 shows the GPS trajectory of our recordings and indicates selective recording sequences along the route. The designed automotive platform will provide a base for future research on pedestrian behaviour analysis.

Acknowledgments

This work was supported by the research project SmartProtect (no. 879642), which is funded through the Austrian Research Promotion Agency (FFG) on behalf of the Austrian Ministry of Climate Action (BMK) via its Mobility of the Future funding program.

References

[1] T. Stewart, "Overview of motor vehicle crashes in 2020 (report no. dot hs 813 266)," *National Highway Traffic Safety Administration*, 2022.

[2] "Road fatalities per billion vehicle kilometers traveled in selected

countries in 2015 [graph]." https://www.statista.com/statistics/485483/road-fatalities-per-billion-vehicle-kilometers-in-selected-countries/, accessed: December 19, 2022.

[3] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *IEEE International Conference on Computer Vision Workshop*, pp. 206–213, 2017.

[4] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *IEEE International Conference on Computer Vision*, pp. 6261–6270, 2019.

[5] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, pp. 3485–3492, 2020.

[6] S. Malla, B. Dariush, and C. Choi, "Titan: Future forecast using action priors," in *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 11186–11196, 2020.

[7] J. Gesnouin, S. Pechberti, B. Stanciulescu, and F. Moutarde, "Assessing cross-dataset generalization of pedestrian crossing predictors," in *IEEE Intelligent Vehicles Symposium*, pp. 419–426, 2022.

[8] K. Man and J. Chahl, "A review of synthetic image data and its use in computer vision," *Journal of Imaging*, vol. 8, no. 11, 2022.

[9] M. Priisalu, C. Paduraru, A. Pirinen, Cristian, and Sminchisescu, "Semantic synthesis of pedestrian locomotion," in *Asian Conference on Computer Vision*, 2020.

[10] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[11] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, and et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2443–2451, 2020.

[12] M. Goebl and G. Farber, "A real-time-capable hard-and software architecture for joint image and knowledge processing in cognitive automobiles," in *IEEE Intelligent Vehicles Symposium*, pp. 734–740, 2007.

[13] E. Andreotti, P. B. Baykas, and S. Selpi, "Mathematical definitions of scene and scenario for analysis of automated driving systems in mixed-traffic simulations," *IEEE Transactions on Intelligent Vehicles*, vol. 6, pp. 366–375, 2021.

[14] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer, "Defining and substantiating the terms scene, situation, and scenario for automated driving," in *IEEE International Conference on Intelligent Transportation Systems*, pp. 982–988, 2015.

[15] X. Qin, S. He, Z. V. Zhang, M. Dehghan, and M. Jägersand, "Bylabel: A boundary based semi-automatic image annotation tool," in *IEEE Workshop on Applications of Computer Vision*, pp. 1804–1813, 2018.

[16] A. Dutta and A. Zisserman, "The vgg image annotator (via)," *ArXiv*, vol. abs/1904.10699, 2019.

[17] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, pp. 157–173, 2008.

[18] B. Pande, K. Padamwar, S. Bhattacharya, S. Roshan, and M. Bhamare, "A review of image annotation tools for object detection," in *International Conference on Applied Artificial Intelligence and Computing*, pp. 976–982, 2022.

[19] F. Groh, D. Schörkhuber, and M. Gelautz, "A tool for semi-automatic ground truth annotation of traffic videos," in *IS&T International Symposium on Electronic Imaging: Autonomous Vehicles and Machines*, pp. 200–1–200–7, 2020.

[20] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *ArXiv*, vol. abs/1805.04687, 2018.

[21] C. Vondrick, D. J. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, vol. 101, pp. 184–204, 2012.

[22] T. A. Biresaw, T. H. Nawaz, J. M. Ferryman, and A. I. Dell, "Vitbat: Video tracking and behavior annotation tool," in *International Conference on Advanced Video and Signal Based Surveillance*, pp. 295–301, 2016.

[23] A. Shen, "Beaverdam : Video annotation tool for computer vision training labels," *EECS Department, University of California, Berkeley, Master Thesis*, 2016.

[24] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, and et al., "Computer vision annotation toolkit (cvat)." https://www.cvat.ai/.

[25] H. Schmidt, J. Terwilliger, D. Aladawy, and A. Fridman, "Hacking nonverbal communication between pedestrians and vehicles in virtual reality," *ArXiv*, vol. abs/1904.01931, 2019.

[26] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *IEEE Intelligent Vehicles Symposium*, pp. 264–269, 2017.

[27] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 8122–8131, 2021.

Author Biography

Dominik Schörkhuber is a PhD student and research assistant at Vienna University of Technology (TU Wien), Austria. He received his bachelor's and master's degree in 2016 and 2018, respectively. His research is centered around video analysis in application to autonomous and assisted driving.

Roman Popp received his PhD in Electrical Electrical Engineering at the TU Wien in 2012. He joined ZKW in 2020 and is mainly working on sensor integration in predevelopment projects.

Oleksandr Chistov received the master's degree in mechatronics/robotics from FH Technikum Wien, Vienna, Austria in 2019 and joined 2020 the ZKW predevelopment team as project manager. He works on multiple domains spanning across machine/deep learning, analytics, image processing and computer vision algorithms with a focus on sensor integration within headlamps for future autonomous driving solutions.

Fabian Windbacher is a deep learning engineer at emotion3D. He received his bachelor's and master's degree from TU Wien in 2021 and 2022, respectively. His research focuses on 3D human pose estimation.

Michael Hödlmoser is CTO at emotion3D and is responsible for all R&D activities within the company. He received his master's degree in Computer Science from Salzburg University of Applied Sciences and his PhD in Visual Computing from TU Wien in 2008 and 2013, respectively. His research focuses on 3D scene understanding and image-based human analysis, with special focus on automotive applications.

Margrit Gelautz is an associate professor at TU Wien, Austria. She received her PhD in Telematics from Graz University of Technology in 1997. Her research focuses on 3D vision, stereo processing, motion analysis and image matting, with special interest in autonomous driving and human-robot interaction.