Providentia – A Large-Scale Sensor System for the Assistance of Autonomous Vehicles and Its Evaluation

Annkathrin Krämmer*, Christoph Schöller*, Dhiraj Gulati, Venkatnarayanan Lakshminarasimhan,
Franz Kurz, Dominik Rosenbaum, Claus Lenz, and Alois Knoll

Abstract—The environmental perception of an autonomous vehicle is limited by its physical sensor ranges and algorithmic performance, as well as by occlusions that degrade its understanding of an ongoing traffic situation. This not only poses a significant threat to safety and limits driving speeds, but it can also lead to inconvenient maneuvers. Intelligent Infrastructure Systems can help to alleviate these problems. An Intelligent Infrastructure System can fill in the gaps in a vehicle's perception and extend its field of view by providing additional detailed information about its surroundings, in the form of a digital model of the current traffic situation, i.e. a digital twin. However, detailed descriptions of such systems and working prototypes demonstrating their feasibility are scarce. In this paper, we propose a hardware and software architecture that enables such a reliable Intelligent Infrastructure System to be built. We have implemented this system in the real world and demonstrate its ability to create an accurate digital twin of an extended highway stretch, thus enhancing an autonomous vehicle's perception beyond the limits of its on-board sensors. Furthermore, we evaluate the digital twin with respect to its spatial accuracy, precision, and recall by using aerial images and earth observation methods for generating ground truth data.

Index Terms—Intelligent infrastructure system, autonomous driving, sensor system, data fusion, digital twin, extended environmental perception, evaluation with aerial images.

I. INTRODUCTION

THE environmental perception and resulting scene and situation understanding of an autonomous vehicle are limited by the available sensor ranges and object detection performance. Even in the vicinity of the vehicle, the existence of occlusions leads to incomplete information about its environment. The resulting uncertainties pose a safety threat not only to the autonomous vehicle itself but also to other road users. To enable it to operate safely, it is necessary to reduce its driving speed, which in turn slows down traffic. This results in

Fig. 1. One of the Providentia measurement points on the A9 highway. The two radars directed towards the north are installed on the other side of the gantry bridge and are therefore not visible from this perspective.

impaired driving comfort, as the vehicle spontaneously reacts to unforeseen scenarios.

Intelligent Infrastructure Systems (IIS) can alleviate these problems by providing autonomous vehicles – as well as conventional vehicles and drivers – with additional information about each road user and the overall traffic situation [1], [2]. In particular, an IIS can observe and detect road users from multiple superior perspectives, with extended coverage compared to that of an individual vehicle. Providing a vehicle with this additional information gives it a better understanding of its surrounding scene and enables it to plan its maneuvers more safely and conveniently. Furthermore, an IIS with the described capabilities enables a multitude of services that further support decision making.

However, actually building such a system involves a number of challenges, such as the right choice of hardware and sensors, and their optimal deployment and utilization in a complex software stack. Its perception must remain reliable and robust in a wide variety of weather, light and traffic conditions. Ensuring such reliability necessitates a combination of multimodal sensors, redundant road coverage with overlapping field of views (FoV), accurate calibration [3], and robust detection and data fusion algorithms.

Since we sketched ideas about how such a system could be designed in previous work [4], in this paper we propose a concrete, scalable architecture. This architecture is the result of experience we made with the real world build-up of the IIS Providentia (see Fig. 1). It includes the system's hardware as well as the software to operate it. In terms of hardware,

we discuss the choice of sensors, the network architecture and the deployment of edge computing devices to enable fast and distributed processing of heavy sensor loads. We outline our software stack and the detection and fusion algorithms used to generate an accurate and consistent model of the world, which we call the digital twin. The digital twin includes information such as position, velocity, vehicle type and a unique identifier for every observed vehicle. By providing this digital twin to an autonomous driving research vehicle, we demonstrate that it can be used to extend the limits of the vehicle's perception beyond its on-board sensors.

For autonomous vehicles to trust the digital twin for maneuver planning, its accuracy and reliability must be known. However, a thorough evaluation requires precise ground truth about the traffic situation on the highway. This is non-trivial to obtain. To solve this issue, we took aerial images of the highway to generate an approximate ground truth and use it to evaluate our system, as explained in [5]. In this paper, we describe in detail the methods used for this evaluation. We present the results of our evaluation of the Providentia system and analyze the system's performance in real world applications. Our evaluation methodology is not specific to our system and can serve as a general framework for the evaluation of IIS.

II. RELATED WORK

First ideas for assisting vehicles, as well as monitoring and controlling traffic with an IIS have already been developed in the *PATH* [6] and *PROMETHEUS* [7] projects in 1986. Recently, with autonomous driving gaining popularity, the need for IIS that are able to support autonomous vehicles has further increased. Several new projects have therefore been initiated, with the goal of developing prototypical IIS. However, their focuses differ widely and few detailed system descriptions are available. The research project *DIGINETPS* [8] focuses in particular on communication topics, considers the detection of pedestrians and cyclists for intersection management, and the communication of traffic signals to vehicles. Similarly, the *Veronika* project [9] focuses on communication between vehicles and traffic signals with the aim of reducing emissions and energy consumption. On the other hand, a system being developed in the *Test Area Autonomous Driving Baden-Württemberg* [10] focuses on providing information for the testing and evaluation of autonomous driving functions by capturing the traffic with multiple sensors. Furthermore, the local highway operator in Austria is transforming its existing road operator system into an IIS [11]. It is also aiming to establish a validation test track and actively supports autonomous vehicles. In the *MEC-View* project, Gabb et al. [12] focus on how data from an IIS can be fused into a vehicle's on-board perception system.

Instead of the IIS itself, many research contributions propose methods of making algorithmic use of the information provided by an IIS, or optimizing their function. With regard to communication networks, Jeffrey Miller [13] proposes an architecture for efficient vehicle-to-vehicle and vehicle-to-infrastructure communication, while Igor Kabashkin [14] analyses the reliability of bidirectional vehicle-to-infrastructure
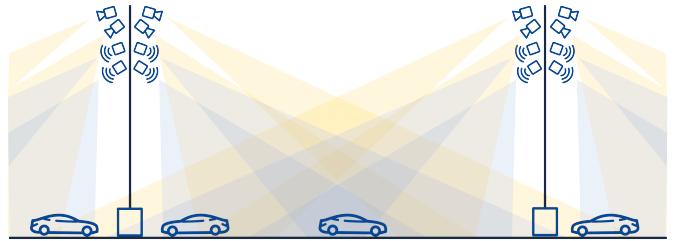

Fig. 2. Schematic illustration of the Providentia sensor setup, with overlapping FoVs for redundancy.

communication. In the project *KoRA9* [15], Geissler et al. [16] formulate an optimization problem that maximizes sensor coverage to locate suitable sensor placements in an IIS. Popular areas of research in the field of computer vision that are related to IIS include traffic density prediction [17], [18] and vehicle re-identification [19], [20]. Other topics involving information provided by an IIS include danger recognition [21] and vehicle motion prediction [22], [23].

Unlike the aforementioned projects and algorithmic literature, this paper focuses on the overall system architecture and implementation of an IIS that generates a digital twin of the highway. The aim of our system is to complete and extend a vehicle's perception and to provide information that enables the implementation of various algorithms and applications based on it. To this end, we conduct a thorough evaluation of our system's performance by considering the overall traffic rather than trajectories from a single test vehicle. Thereby we account for a broad variety of vehicle types, colors and driving behaviors. In particular, we evaluate the spatial accuracy as well as the detection rate, i.e. the system's performance with respect to missing vehicles and false detections.

III. PROVIDENTIA SYSTEM ARCHITECTURE

In this section, we describe the design of the Providentia system, including the hardware and software setup and the algorithms used for detection, calibration and fusion.

Providentia is a large-scale distributed sensor system consisting of multimodal sensors, multiple edge computing units, a complex software architecture, and a broad range of state-of-the-art algorithms. It is built along the A9 highway close to Munich. Its primary purpose is to provide a real-time and reliable digital twin of the current road traffic at any given time or day of the year, for use in a variety of applications.

A. Hardware and Software Setup

At the time of writing, two gantry bridges – separated by a distance of approximately $440 \, \mathrm{m}$ – have been equipped with sensors and computing hardware. Each of these gantry bridges represents one measurement point in our system as illustrated in Fig. 1. To achieve high perception robustness, we use sensors of different measurement modalities that cover the entire stretch between our measurement points redundantly. Fig. 2 illustrates the overall setting of the system with the redundant coverage of the highway.

Each measurement point comprises eight sensors with two cameras and two radars per viewing direction. In each direction, one radar covers the right-hand side while the other
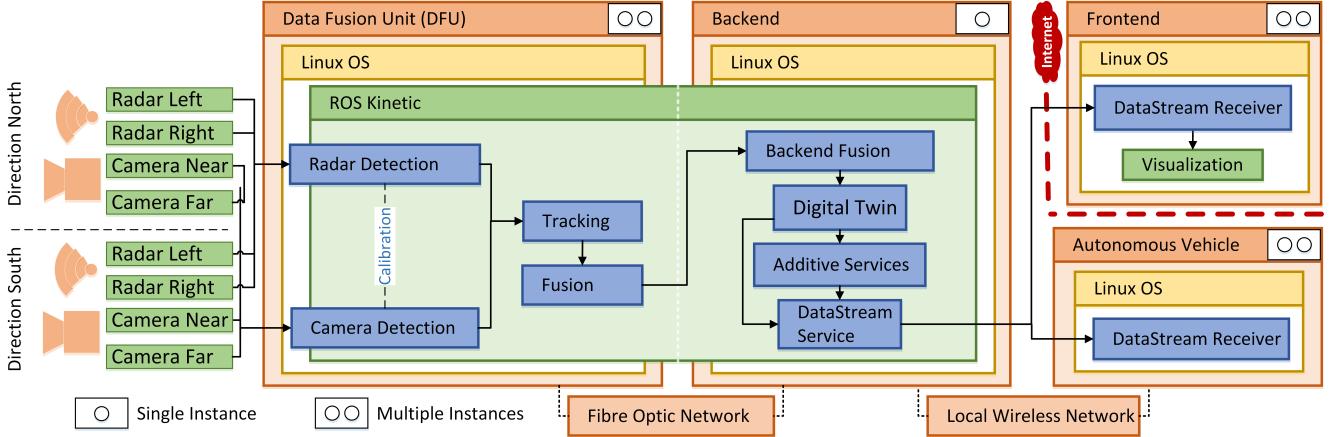
Fig. 3. Platform architecture of the Providentia system.

covers the left-hand side of the highway. The cameras have focal lengths of $16\,\mathrm{mm}$ and $50\,\mathrm{mm}$ to enable them to capture both the far and near ranges. By combining sensors with different measuring principles, our system is able to operate in varying traffic, light and weather conditions. Besides having redundant coverage with the sensors on each measurement point, we also selected the positions of the two measurement points in such a way that their overall FoVs overlap. This further increases redundancy and thus robustness, and allows smooth transitions while tracking vehicles as they move from one measurement point to the other. In addition, covering the highway stretch from different viewing directions helps to resolve detection errors and occlusions.

The system employs specialized $24\,\mathrm{GHz}$ traffic monitoring radars made by SmartMicro, of the generation UMRR-0C, with a type 40 antenna. The cameras are Basler acA1920-50gc. All the sensors at a single measurement point are connected to a Data Fusion Unit (DFU), which serves as a local edge computing unit and runs with Ubuntu 16.04 Server. It is equipped with two INTEL Xeon E5-2630v4 $2.2\,\mathrm{GHz}$ CPUs with $64\,\mathrm{GB}$ RAM and two NVIDIA Tesla V100 SXM2 GPUs. All sensor measurements from the cameras and radars are fed into the detection and data fusion toolchain running on this edge computing unit. This results in object lists containing all the road users tracked in the FoV of that measurement point. Each DFU transmits this object list to a backend machine via a fibre optic network, where they are finally fused into the digital twin that covers the entire observed highway stretch.

The full architecture is shown in Fig. 3. We use ROS on all computing units to ensure seamless connectivity. The final digital twin is communicated either to autonomous vehicles or to a frontend, where it can be visualized as required for drivers or operators.

B. Object Detection

The first step towards creating the digital twin of the highway is to detect and classify the vehicles on the highway. Our traffic radars are capable ex-works of detecting objects and providing time-stamped positions and velocities in their respective local sensor coordinate systems on street level.

We transform this output of each radar into our system's global Cartesian coordinate system using the radar calibration parameters in preparation for data fusion.

Detection and classification of objects in the camera images are performed by the DFU edge devices next to the highway. The system's cameras publish time-stamped images that are tagged with a unique camera identifier. To ensure scalability and safety even in the event of camera failures, our modular object detection pipelines subscribe to each image stream separately. The object detection pipelines are constantly monitored to supervise and analyze detection performance. The modular services are automatically restarted in the event of any failures. Not only are the multiple camera streams processed in parallel, the object detection can also work with various detection networks. This allows us to configure the object detection to optimally balance between low computation time and high accuracy, depending on the requirements that the application of our system poses. To this end, we performed extensive research on state-of-the-art detection algorithms, based on neural networks [24].

At the time of writing, we have been using the YOLOv3 [25] architecture as our detection network in the object detection pipelines. In addition to regressing two-dimensional bounding boxes with a confidence score, this network classifies the detected vehicles into types as car, truck, bus or motorcycle. The output is then published prior to transformation.

To compute the three-dimensional positions of the vehicles from the camera detections in the images, we shoot a ray through the lower-edge midpoint of the bounding box and intersect it with the street-level ground plane that we know from our camera calibration. We transform the resulting vehicle detections in the same manner as the detections of the radars into our system's global Cartesian coordinate system. All of the resulting measurements are then ready to be fused into a consistent world model and are fed into the data fusion pipeline, starting with a tracking module.

C. Calibration

Precise calibration of each sensor and measurement point is necessary to make all of the transformations possible. While

we intrinsically calibrated the cameras individually with the common checkerboard method prior to installation, the overall extrinsic calibration of the system is non-trivial. Not only does our system possess a high number of sensors and degrees of freedom, but it also makes use of sensors with heterogeneous measurement principles. To obtain an approximate starting point for our calibration algorithms, we measured the positions of all sensors using a modern laser distance meter. We determined the sensor angles for our cameras with a digital angle finder and spirit level, while for the radars we used their built-in calibration algorithms. We then refined these estimates for the cameras using vanishing point methods [26]. To fine-tune the inter-sensor calibration, we manually minimized the projection and re-projection errors for all sensor pairs, both in the images and on road level in the three-dimensional global coordinate system. To calibrate the system as a whole with respect to the world (GPS coordinates), we used a GPS receiver and localized the measurement points within a high definition map.

D. Data Fusion

When it comes to the sensor data fusion, a large-scale system such as ours poses many challenges. On the highway, we can observe a very large number of vehicles that have to be tracked in real time. Therefore, the data fusion system has to scale for hundreds of vehicles. In addition, the number of targets is not known in advance and our fusion must be robust with respect to clutter and detection failures. Conventional filtering methods that handle each observed vehicle separately, such as multiple Kalman filters or multiple hypotheses tracking [27], require to explicitly solve a complex association problem between the system's sensor detections and tracked vehicles. This severely limits scalability. For this reason, we use the random finite set (RFS) framework [28], [29], specifically the Gaussian mixture probability hypothesis density (GM-PHD) filter [30]. This filter avoids the explicit data association step and has proven to balance our run-time and scalability constraints well. Additionally, it handles time-varying target numbers, clutter and detection uncertainty within the filtering recursion.

We add tracking capabilities to our GM-PHD filter by extending it with ideas taken from Panta et al. [31]. In particular, we make use of the tree structure that naturally arises in the GM-PHD filter recursion and appropriate track management methods. With the resulting tracker we track the measurements of each sensor in parallel. For motion and sensor models, we use a standard constant velocity kinematic model [32] and a zero-mean Gaussian white noise observation model, respectively. All parameters for our sensor and scenario specifications were tuned empirically. To fuse the tracked data from different sensors and measurement points, we adapt the method from Vasic et al. [33] that is based on generalized covariance intersection [34]. In order to ensure scalability and easy extension of our system setup, we implement a hierarchical data fusion concept, in which we first perform independent local sensor fusion at each measurement point leading to vehicle tracklets. Second-level fusion of all measurement points is then performed on the backend. This step
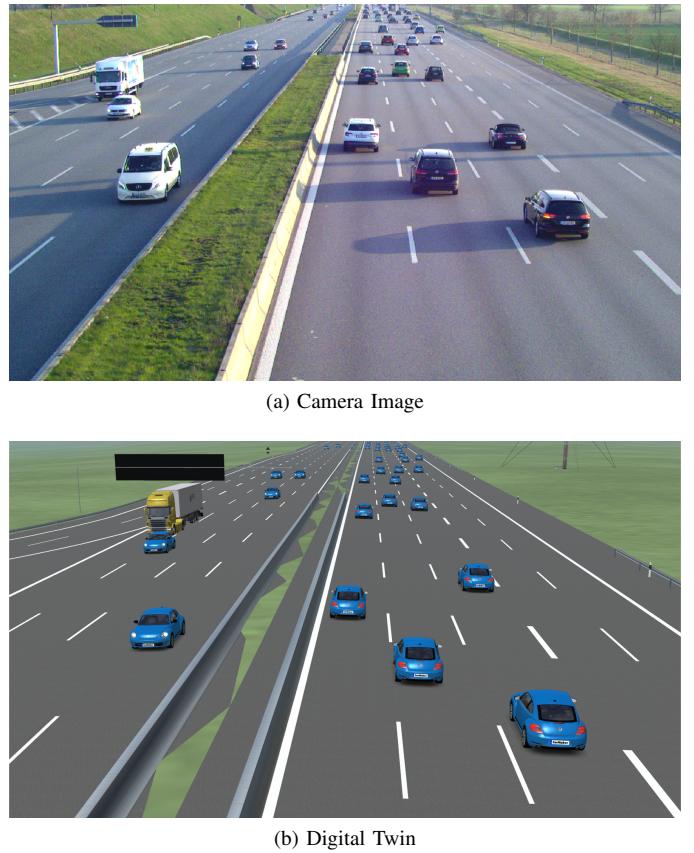

(a) Camera Image

(b) Digital Twin

Fig. 4. Qualitative example of how our system captures the real world (a) in a digital twin (b). We recreate the scene with generalized models for different vehicle types for visualization purposes. During operation, all information is sent to the autonomous vehicle in form of a sparse object list.

generates the consistent model of the whole highway scene covered by our system, that we refer to as the digital twin. It is transformed to GPS coordinates depending on the application it is communicated to.

IV. DIGITAL TWIN AND EXTENSION OF VEHICULAR PERCEPTION

The digital twin represents the main output of the Providentia system. It consists of the position, velocity, and type of every vehicle observed, with each one assigned a unique tracking identifier. It can be used by vehicles on the highway stretch to improve their decision making and to implement additional services that can be provided by the infrastructure itself. Such services might include motion prediction for each vehicle, congestion recognition, lane recommendations, and collision warnings. In this section, we illustrate our system's ability to capture highway traffic and demonstrate its potential for extending an autonomous vehicle's perception of the scene.

Our qualitative examples were captured in our testbed under real-world conditions. Currently, our system redundantly covers a stretch of about $440\,\mathrm{m}$ of the road, which corresponds to the distance between the two measurement points. Fig. 4 shows an example of a digital twin of current traffic on the highway as computed by our system. It is a visualization of the information that is also sent to autonomous vehicles to extend their perception. Our system is able to reliably detect the vehicles passing through the testbed. This is only possible
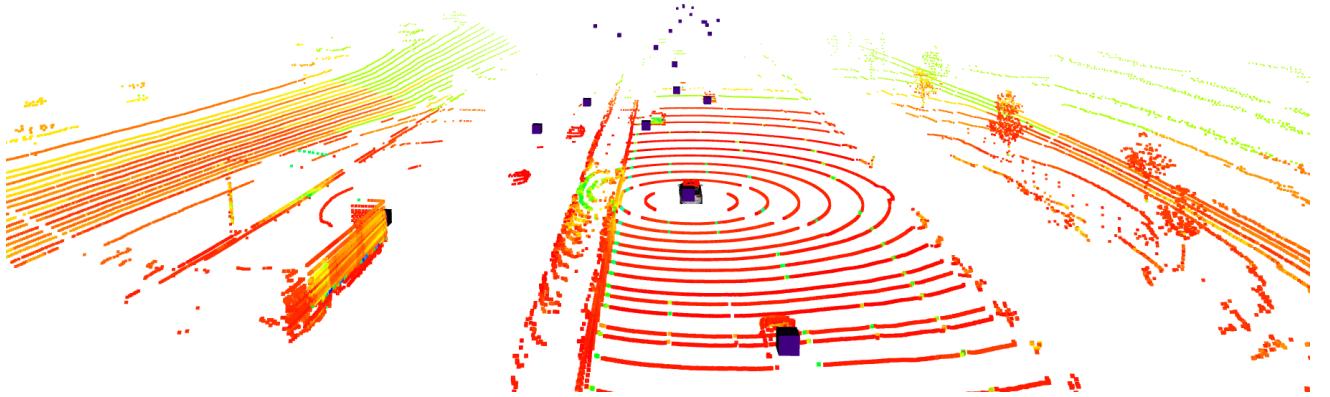
Fig. 5. An autonomous vehicle driving through our testbed. The dots visualize the vehicle's lidar measurements and the purple cubes represent the vehicles perceived by the Providentia system. While the vehicle's own lidar range is severely limited, its perception and resulting scene understanding are extended into the far distance using information from our system.

by fusing multiple sensor perspectives. The update rate for the digital twin depends on the type of the used object detection network and varies between approximately $5\,\text{Hz}$ and $15\,\text{Hz}$. As mentioned in Sec. III, this represents a trade-off between detection speed and accuracy. As our aim in this study is to generate an accurate and reliable digital twin, we use a network version with high quality detections and achieve a frequency of $5.4\,\text{Hz}$.

We also transmitted this digital twin to our autonomous driving research vehicle *fortuna* [35] for the purpose of extending its environmental perception and situation understanding. Vehicles perceive their environment by means of lidars, which have limited measurement ranges and the point cloud density in the distance becomes increasingly sparse. Vehicular cameras can capture a more distant environment than lidars are able to, but objects that are too far away appear small on the image and cannot be reliably detected. Furthermore, the vehicle's low perspective is prone to severe occlusions. Fig. 5 illustrates how an autonomous vehicle driving through our system perceives its environment. The violet cubes represent vehicles detected by our system. We observed that the point cloud density of our vehicle's lidars drops significantly at a distance of approximately $80\,\text{m}$, but our system's digital twin extends the vehicle's environmental perception to up to $400\,\text{m}$. In principle, a system such as ours is able to extend the perception of a vehicle even further, since we designed it with scalability in mind. The maximum distance is only limited by the existing number of built-up measurement points.

V. Evaluation of the Digital Twin

To decide what applications can be realized with our system, it is crucial to know the spatial accuracy and detection rate of the digital twin that it generates. For example, using the digital twin for maneuver planning in autonomous vehicles requires high position accuracy, whereas position accuracy is less important for the detection of traffic jams. Knowing the statistical certainty and uncertainty of the system's measurements also makes it possible to define safety margins that vehicles have to take into account when using the provided information.

However, as we explained in [5], the evaluation of the system's digital twin is a challenging task. To merely evaluate the detection performance of individual sensors is insufficient to judge the system's performance, as the calibration between the sensors and the fusion algorithms are of paramount importance for the quality of the digital twin and must therefore also be included in the evaluation. End-to-end evaluation of the system requires ground truth information of the traffic on the testbed over an extended period of time. This implies having the exact positions of all the vehicles on the observed stretch of highway. Labeling the images from the cameras within the system is not sufficient, as it would only provide ground truth information in image coordinates but not in the real world. Neither is using a single, localized test vehicle sufficient, as the system must be able to handle a wide variety of vehicle colors and shapes. Furthermore, the usefulness of simulations is also limited. In reality, the system is subject not only to various lighting and vibration effects, but also to the decisions of drivers, which are hard to model.

One way of approximating the required ground truth is to record aerial images of the testbed. These have an ideal – almost orthogonal – top-down perspective of the highway. This perspective avoids all inter-vehicle occlusions, and due to their regular contours, vehicles are easy to detect and distinguish. In this section, we will describe how we captured and processed these images to generate ground truth data that is suitable for evaluating our system. We also explain the evaluation itself in detail and discuss the results together with their implications for the performance of our system.

A. Ground Truth Generation

As previously outlined in [5], to generate such an aerial view ground truth, the Providentia testbed was recorded using a 4k camera system mounted on a H135 helicopter. Both the camera system and the Providentia system were synchronized with GPS time. The 4k camera system consists of three Canon EOS 1D-X Mark II cameras oriented in different viewing directions, each recording images at a resolution of 20.2 megapixels. The cameras to the left and right covered the northern and southern
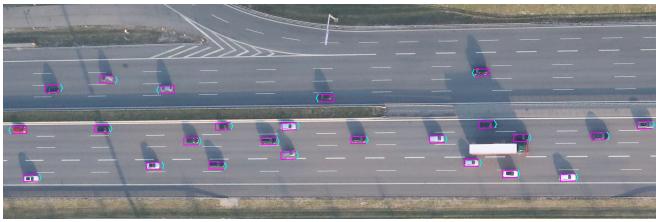
Fig. 6. Crop of an aerial image including vehicle detections, taken with the helicopter's left-hand camera that captures part of our testbed. Note that the current system only detects cars and not trucks.
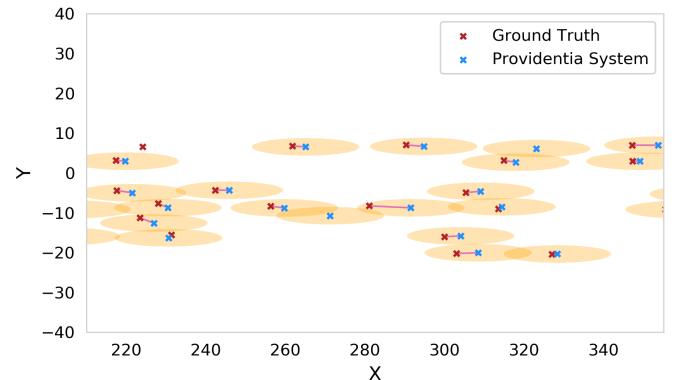

Fig. 7. Associations between detections from our system and ground truth data. Only ground truth detections within the displayed ellipse are associated and actual associations are marked with a line between corresponding detections.

parts of the testbed respectively with an overlapping FoV. The third, nadir-looking camera of the system was not used. The cameras captured images simultaneously at a rate of one image per second at a flight altitude of 450 m above ground, covering an area of 600 m × 250 m. With a focal length of 50 mm, each image pixel corresponds to 6 cm on the ground. In addition, an IGI Compact MEMS GNSS/IMU system was used to estimate the position and orientation of the sensors during flight to enable georeferencing of the images captured. To optimize the georeferencing accuracy, bundle adjustment with tie points and ground control points was performed.

We use a neural network that has been trained and evaluated with the DLR-MVDA dataset [36] for object detection in all aerial images. Due to the lack of trucks in the training data, it is only able to detect cars. This is why we focus on this object type within the digital twin in our evaluation. To compute the positions of the detected vehicles on the road, we cast rays through all four bounding box corner positions in the aerial image and intersect them with a lidar terrain model of the highway surface to compute local UTM coordinates. To obtain the final ground truth data that we use in our evaluation, we compute the center position of each vehicle. We filter out all vehicles that are detected outside of the Providentia system's field of view or are clearly erroneous, such as vehicle detections oriented perpendicularly to the driving direction and vehicles detected twice in the overlap of the two camera FoVs.

The quality of the obtained ground truth depends on the accuracy of the object detections in the aerial images. The detection quality is the subject of on-going research [37]. Our ground truth dataset rarely contained false or severely misplaced detections and only a few dark-colored vehicles were missed. The position accuracy of the detected vehicles depends on the georeferencing accuracy of the images. Specifically, it depends on the calibration accuracy of the 4k system, the quality of the tie and ground control points used for bundle adjustment, and the accuracy of the underlying terrain model. The overall absolute accuracy on the present dataset lies in the centimeter range. The accuracy of this georeferencing is demonstrated in [38].

Fig. 6 shows a captured aerial image with vehicle detections. Even though the testbed was not always fully covered by the helicopters' cameras, we captured enough vehicles to perform a reliable and statistically significant evaluation. In total, we generated 2 minutes of ground truth data containing 2139 valid vehicle observations.

B. Evaluation Methodology

To compare our digital twin with the ground truth data, we first transform the aerial detections from UTM into the Cartesian world coordinate system used by Providentia in which the digital twin is defined. In the next step, we match all ground truth frames with those frames of the digital twin with the closest corresponding timestamp. Note that the digital twin has a higher frequency of 5.4 Hz than the ground truth with 1 Hz, and they are slightly offset with respect to each other. We account for the time difference between matched frames by extrapolating the Providentia detections with a constant velocity prediction. This is sufficiently accurate, as the average time difference is only 50 ms.

In order to evaluate the spatial accuracy and the detection rate of the digital twin, it is necessary to establish associations between the detected vehicles in the digital twin and the ground truth. For this purpose we use the Hungarian algorithm [39], which guarantees an optimal one-to-one assignment. A weighted Euclidean distance is used for computing the cost matrix, putting more emphasis on the longitudinal distance between vehicles than on their lateral distance. This accounts for the fact that the estimates in the digital twin have a greater variance in the driving direction (see the RMSE results in Sec. V-C), while vehicles driving on nearby lanes can be close, but should not be associated. We chose the weights empirically, such that they represent a 13.5 m by 2.2 m ellipse around the ground truth object, with the major axis aligned with the driving direction (see Fig. 7). Based on this ellipse, we also perform a gating, in which all associations outside of the ellipse are rejected. Overall, these parameters lead to accurate and reasonable associations.

With these associations established, we evaluate the spatial accuracy of the digital twin by computing the root-mean-square error (RMSE). It represents the standard deviation of the metric error between the vehicle positions in the digital twin and the ground truth positions. In addition to spatial accuracy, the detection rate of our system is important for evaluating its overall performance. Appropriate metrics for this are precision and recall. The precision of our system states which percentage of vehicles detected by our system
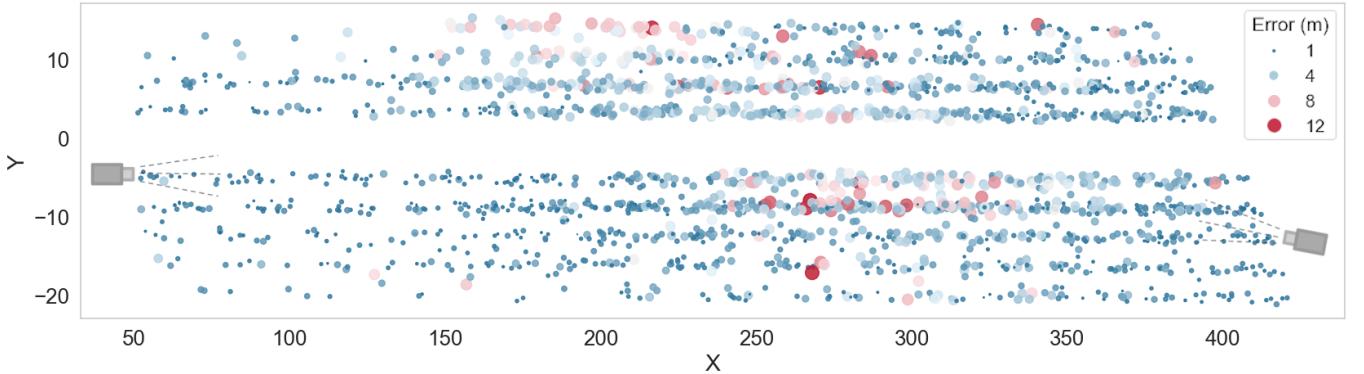
Fig. 8. Positioning errors in the digital twin for each ground truth vehicle on the highway. The camera icons represent the measurement point positions. Severe errors are rare and are mostly due to unfavorable camera perspectives (on the top lane) or a great distance to the sensor (towards the middle of the testbed).
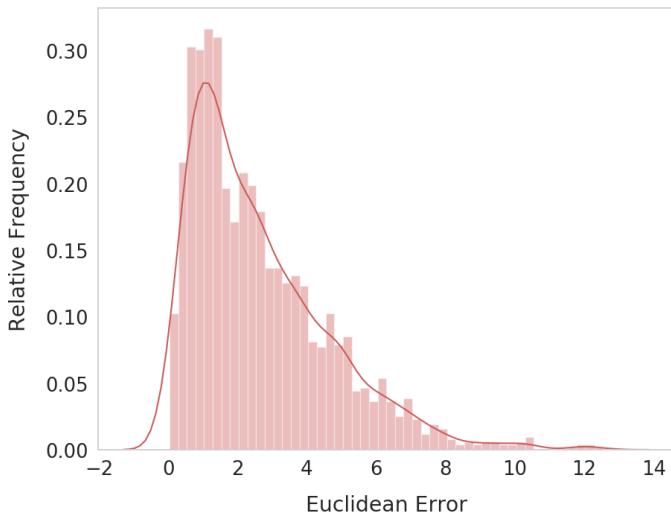

Fig. 9. Distribution of positioning errors in the digital twin compared to the aerial ground truth. Our system is very accurate in most cases. In 50 % of cases the errors are less than 2.16 m, i.e. lie within the vehicle's extent, and in 95 % of cases the errors are less than 6.55 m.

were actually present. Its recall refers to the percentage of vehicles in the testbed successfully detected by our system. To evaluate these metrics, we projected the detections from our system onto the aerial camera images and manually counted the true positives, false positives and false negatives. Automatic computation of false detections would result in an inaccurate false positive count, as the helicopter was moving and occasionally only covered parts of our testbed with its FoV.

The evaluation of our system was performed during the day with a medium traffic volume. Furthermore, we only used the cameras of our system, because during the day these are more accurate than our installed radars in terms of their ability to estimate vehicle positions. In our evaluation, we considered the area enclosed by the two measurement points that we cover redundantly.

C. Results

Our results for the spatial accuracy and detection rate of the digital twin are summarized in Table I. By evaluating the positioning accuracy of our system, we achieve an RMSE of 3.31 m. Of this error, the longitudinal direction makes up the major part with 3.27 m, while the RMSE in the lateral direction is only 0.53 m. This high lateral accuracy allows us to reliably determine the lane for each vehicle.

The major part of both positioning errors can be explained by our current lack of information regarding the extents of the objects. Because our camera detections are two-dimensional bounding boxes in the image plane, the vehicles' lengths are not taken into account for estimating their position, and their widths can only be approximated because of perspective errors. Therefore, the detection positions in our system are generally placed either at the rears or at the fronts of the vehicles, depending on the perspective. Hence, the position estimate in the digital twin varies over the vehicle extent within the testbed. However, in the ground truth, the position of an object is specified at its center. Taking this into account, our spatial accuracy is very promising. As the ground truth cars driving through the testbed over the course of our evaluation have an average length of approximately 4.6 m, the placement of a detection at the rear or front would already cause a displacement of about 2.3 m to the center. Additionally, Fig. 9 shows the distribution of our positioning errors. In 50 % of cases, our error is less than 2.16 m, i.e. the digital twin detections lie within the car's extent, while in 95 % of cases, it is less than 6.55 m. This indicates that by incorporating the vehicles' extents from the sensors, e.g. by computing three-dimensional bounding boxes, this error could be greatly reduced.

Fig. 8 shows how the positioning errors are distributed over our testbed. Larger positioning errors are relatively rare and mainly occur towards the middle of the testbed. This is the

area in which the vehicles are farthest away from all the sensors. This means that the resolution of the vehicles in the image is the smallest, which results in higher uncertainty in the bounding box estimates. Also, the way we compute the vehicle positions from the camera detections, i.e. by intersecting rays with the road (see Sec. III-B), is more sensitive to inaccuracies over large distances. Furthermore, as the road is only approximately planar, deviations from this plane have a greater influence at the distance because the rays intersect the ground plane at a flatter angle. Additionally, our sensors are exposed to vibrations that make them susceptible to calibration inaccuracies. Angular calibration inaccuracies in particular have a greater influence the greater the distance.

Based on the distribution of errors, it is also evident that the left-hand measurement point has an advantageous perspective of the road, as compared to the right-hand measurement point. Its sensors are positioned more centrally on the highway and are better aligned with the driving direction of the vehicles. The right-hand measurement point has a more oblique perspective of the vehicles, resulting in the lower-edge midpoints of the bounding boxes deviating more from the actual middle of the vehicles. For this reason, the errors in the middle of the testbed tend to be larger towards the right-hand measurement point. Moreover, it is because of the measurement points' perspectives that the top lane is the most error-prone. The sensors of both measurement points are installed on gantry bridges over the opposite side of the highway. Despite these perspective errors, our overall positioning accuracy is high, especially closer to the measurement points.

As for the detection rate, our system achieves a precision of 99.38%. This means that almost all cars detected by our system do actually exist and are correct, meaning that we have very few false positives. Our system achieves a recall of 97.43%. This means that we detect 97.43% of the ground truth cars on the highway and miss only 2.57% of them. However, most of the false negatives – about 55% – occur at the boundary of the Providentia system's overall FoV. One reason for this is that it is difficult to determine an exact FoV boundary because of the vibrations that our sensors are subject to. This means that there are cars in the ground truth that we did not filter out but that are just slightly outside of the current actual FoV and cannot be detected by our system. These are not actual false negatives. The second reason is that the object detector might not detect a vehicle when it first appears on the camera image due to it being only captured partially. And once detected, the birth model of the tracker only initializes tracks from two consecutive detections to reduce noise (thus the few false positives). Therefore, the tracker needs an additional timestep before establishing the car as an estimate in the digital twin.

With only about 45% of the false negatives occurring within the testbed, this leads to an additional increase of approximately 1.4% in recall, when excluding the boundaries. This can be considered to be the actual recall of our system, since detections do not need to be communicated to autonomous vehicles during the initialization phase.

It is important to note that we analyzed precision and recall on a frame-by-frame basis. Hence, when we do not detect a vehicle, it does not imply that it is passing through the testbed completely undetected. This is very unlikely to happen. Rather, at specific moments in time, certain vehicles may be briefly lost. This mostly occurs as described above in specific areas, or due to occlusions caused by large vehicles such as trucks.

Overall, our system achieves a high degree of reliability, both in terms of positioning accuracy and detection rate.

VI. Conclusion

To improve the safety and comfort of autonomous vehicles, one should not rely solely on on-board sensors, but their perception and scene understanding should be extended by adding information available from a modern IIS. With its superior sensor perspectives and spatial distribution, an IIS can provide information far beyond the perception range of an individual vehicle. This can resolve occlusions and lead to better long-term planning of the vehicle.

While there is much research currently being done on specific components and use-cases of IIS, information on building up an entire system is sparse. This paper describes how a modern IIS can be successfully designed and built. This includes the hardware and sensor setup, detection algorithms, calibration, and data fusion. With our evaluation we have shown that our system is able to achieve reasonable results at capturing traffic on an observed highway stretch and that it can generate a reliable digital twin in near real time. We have further demonstrated that it is possible to integrate the information captured by our system into the environmental model of an autonomous vehicle to extend its limited perception range.

Our extensive quantitative evaluation has shown that our system is characterized by both a high detection rate and spatial accuracy. The primary purpose of our system is to enhance the perception of autonomous vehicles in the testbed. But based on the results of our evaluation, it is also evident that a such a system could be used for applications such as traffic prediction or the detection of emerging traffic jams, wrong-way drivers and immobile vehicles. Traffic flow management with lane and speed recommendations could be another possible application. Beyond this, the system could be used as a reference for testing, evaluating and developing autonomous driving functions.

Both in general and taking into consideration the above, our system could be improved in terms of the update rate of its digital twin, its night-time functionality and its operation in adverse weather conditions, as well as during traffic jams with severe occlusions. Taking into account the vehicles' spatial extents would improve its positioning accuracy and reduce errors caused by different camera perspectives. Furthermore, methods of automatically calibrating the system as a whole would greatly improve its performance and reliability. Vibrations, oscillations of measurement points and temperature changes can lead to deterioration even of accurately calibrated systems if they are not continuously adjusted.

References

[1] K. N. Qureshi and A. H. Abdullah, "A survey on intelligent transportation systems," *Middle-East J. Sci. Res. (MEJSR)*, vol. 15, no. 5, pp. 629–642, 2013.

[2] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer, "Uav-enabled intelligent transportation systems for the smart city: Applications and challenges," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 22–28, 2017.

[3] C. Schöller, M. Schnettler, A. Krämmer, G. Hinz, M. Bakovic, M. Güzet, and A. Knoll, "Targetless rotational auto-calibration of radar and camera for intelligent transportation systems," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, 2019, pp. 3934–3941.

[4] G. Hinz, M. Büchel, F. Diehl, G. Chen, A. Krämmer, J. Kuhn, V. Lakshminarasimhan, M. Schellmann, U. Baumgarten, and A. Knoll, "Designing a far-reaching view for highway traffic scenarios with 5g-based intelligent infrastructure," in *8. Tag. Fahrerassist.*, 2017.

[5] A. Krämmer, C. Schöller, F. Kurz, D. Rosenbaum, and A. Knoll, "Vorausschauende Wahrnehmung für sicheres automatisiertes Fahren. Validierung intelligenter Infrastruktursysteme am Beispiel von Providentia," *Int. Verkehrsw.*, vol. 72, no. 1, pp. 26–31, 2020.

[6] S. E. Shladover, "The california path program of ivhs research and its approach to vehicle-highway automation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, 1992, pp. 347–352.

[7] H.-H. Braess and G. Reichart, "Prometheus: A vision of the intelligent car on intelligent roads? an attempted critical appraisal. I," *Automobiltechn. Zeitschr. (ATZ)*, vol. 97, no. 4, pp. 200–205, 1995.

[8] Diginetps - the digitally connected protocol track. Accessed: June 8, 2020. [Online]. Available: https://diginet-ps.de/en/home/

[9] Veronika. Accessed: June 8, 2020. [Online]. Available: https://www.bmvi.de/SharedDocs/DE/Artikel/DG/AVF-projekte/veronika.html

[10] Test area autonomous driving Baden-Württemberg – field test for research and innovation. Accessed: June 8, 2020. [Online]. Available: https://taf-bw.de/en/

[11] S. Seebacher, B. Datler, J. Erhart, G. Greiner, M. Harrer, P. Hrassnig, A. Präsent, C. Schwarzl, and M. Ullrich, "Infrastructure data fusion for validation and future enhancements of autonomous vehicles' perception on austrian motorways," in *Proc. IEEE Int. Conf. Connec. Vehicles Expo (ICCVE)*, 2019, pp. 1–7.

[12] M. Gabb, H. Digel, T. Müller, and R.-W. Henn, "Infrastructure-supported perception and track-level fusion using edge computing," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, 2019, pp. 1739–1745.

[13] J. Miller, "Vehicle-to-vehicle-to-infrastructure (v2v2i) intelligent transportation system architecture," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, 2008, pp. 715–720.

[14] I. Kabashkin, "Reliability of bidirectional v2x communications in the intelligent transport systems," in *Adv. Wireless Opt. Commun. (RTUWO)*, 2015, pp. 159–163.

[15] Kooperative Radarsensoren für das digitale Testfeld A9 - KoRA9. Accessed: July 10, 2020. [Online]. Available: https://www.bmvi.de/SharedDocs/DE/Artikel/DG/AVF-projekte/KoRA9.html

[16] F. Geissler and R. Gräfe, "Optimized sensor placement for dependable roadside infrastructures," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, 2019, pp. 2408–2413.

[17] S. Zhang, G. Wu, J. P. Costeira, and J. M. F. Moura, "Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras," in *Int. Conf. Comput. Vision (ICCV)*, 2017, pp. 3667–3676.

[18] ——, "Understanding traffic density from large-scale web camera data," in *Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2017, pp. 4264–4273.

[19] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *Int. Conf. Comput. Vision (ICCV)*, 2017, pp. 1918–1927.

[20] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 6489–6498.

[21] L. Yu, D. Zhang, X. Chen, and A. Hauptmann, "Traffic danger recognition with surveillance cameras without training data," in *Int. Conf. Adv. Video Signal Surveill. (AVSS)*, 2018, pp. 378–383.

[22] F. Diehl, T. Brunner, M. Truong Le, and A. Knoll, "Graph neural networks for modelling traffic participant interaction," 2019. [Online]. Available: https://arxiv.org/abs/1903.01254v2

[23] J. Liu, Y. Luo, H. Xiong, T. Wang, H. Huang, and Z. Zhong, "An integrated approach to probabilistic vehicle trajectory prediction via driver characteristic and intention estimation," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, 2019, pp. 3526–3532.

[24] F. Altenberger and C. Lenz, "A non-technical survey on deep convolutional neural network architectures," 2018. [Online]. Available: https://arxiv.org/abs/1803.02129

[25] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018. [Online]. Available: https://arxiv.org/abs/1804.02767

[26] N. K. Kanhere and S. T. Birchfield, "A taxonomy and analysis of camera calibration methods for traffic monitoring applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 441–452, 2010.

[27] S. S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 19, no. 1, pp. 5–18, 2004.

[28] R. P. S. Mahler, *Statistical multisource-multitarget information fusion*. Artech House, 2007.

[29] ——, *Advances in statistical multisource-multitarget information fusion*. Artech House, 2014.

[30] B.-N. Vo and W.-K. Ma, "The gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, 2006.

[31] K. Panta, D. E. Clark, and B.-N. Vo, "Data association and track management for the gaussian mixture probability hypothesis density filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 45, no. 3, pp. 1003–1016, 2009.

[32] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll, "What the constant velocity model can teach us about pedestrian motion prediction," *Robot. Autom. Lett. (RA-L)*, vol. 5, no. 2, pp. 1696–1703, 2020.

[33] M. Vasic and A. Martinoli, "A collaborative sensor fusion algorithm for multi-object tracking using a gaussian mixture probability hypothesis density filter," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, 2015, pp. 491–498.

[34] R. P. S. Mahler, "Optimal/robust distributed data fusion: A unified approach," in *Signal Process., Sensor Fusion, Target Recognit. IX*, vol. 4052, 2000, pp. 128–138.

[35] T. Kessler, J. Bernhard, M. Büchel, K. Esterle, P. Hart, D. Malovetz, M. Truong Le, F. Diehl, T. Brunner, and A. Knoll, "Bridging the gap between open source software and vehicle hardware for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, 2019, pp. 1612–1619.

[36] Dlr dataset for multi-class vehicle detection and orientation in aerial imagery (DLR-MVDA). Accessed: June 8, 2020. [Online]. Available: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-12760/22294_read-52777/

[37] S. M. Azimi, C. Henry, L. Sommer, A. Schumann, and E. Vig, "Skyscapes – fine-grained semantic understanding of aerial scenes," in *Int. Conf. Comput. Vision (ICCV)*, 2019, pp. 7393–7403.

[38] F. Kurz, T. Krauß, H. Runge, D. Rosenbaum, and P. Angelo, "Precise aerial image orientation using sar ground control points for mapping of urban landmarks," in *Int. Arch. Photogram., Remote Sens., Spatial Inf. Sci. (ISPRS Archives)*, vol. 42, no. 2, 2019, pp. 61–66.

[39] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.