Deliverable

WP3 – Development of short-range 3D-imaging systems

D3.11 Annotated benchmarking data base for in-cabin scenarios

Project Information

Grant Agreement n°	826600
Dates	01/05/19 - 31/10/22

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826600. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Sweden, Greece, Spain, United Kingdom, Germany, Luxembourg, Latvia, Hungary.
The VIZTA project results presented reflect only the author's view. The JU is not responsible for any use that may be made of the information it contains.

Document status

Document Information

Deliverable name	VIZTA_D3.11_07042021_VF
Responsible beneficiary	DFKI
Contributing beneficiaries	IEE S.A.
Contractual delivery date	M22 – 28/02/21
Actual delivery date	M24 – 07/04/21
Dissemination level	Public

Document approval

Name	Position in project	Organization	Date	Visa
Karine Perin	WP3 Leader	ST GNB2 SAS	22/02/21	OK
Laurent Dugoujon	Coordinator	ST SAS C2	17/03/21	OK
Maud Bossard	MST member	Ayming	07/04/21	OK

Document history

Version	Date	Modifications	Authors / Organization
V1	19/01/21	First version	J. Katrolia, B. Mirbach, J.Rambach / DFKI, F. Grandidier /IEE
V2	02/02/21	Review	K. Perin / ST GNB2 SAS
V3	15/02/21	Revision	B. Mirbach / DFKI
V4	22/02/21	Review of complementary information	K. Perin / ST GNB2 SAS
V5	17/03/21	Review by Coordinator	L.Dugoujon ST GNB2 SAS
V6	23/03/21	Links added	B. Mirbach / DFKI
VF	07/04/21	Coordinator's approval	L.Dugoujon ST GNB2 SAS

Table of content

Executive summary

This deliverable describes an extensive dataset of annotated real time-of-flight (ToF) data recorded at the [DFKI] in-cabin test platform (deliverable D3.31) as well as complementary synthetic data generated using rendering of 3D car model. The dataset covers several subjects with a variation of equipment and objects such as child seats to generate a realistic subset of real-world in-car cabin scenarios. The data is annotated with 2D and 3D bounding boxes and segmentation masks. The dataset of in-cabin scenes allowed the development of detection and segmentation algorithms for the AS2 demonstrator (sub-task 3.C.3), described in deliverable D3.34. [DFKI] plans to make the dataset openly available to the scientific community after publication of the corresponding conference paper.

1 Brief description of the state of the art

With the rise of deep learning-based applications to in-cabin monitoring, several datasets have emerged in parallel to complement these data hungry methods. In many cases these datasets encompass several modalities of data like RGB, infrared gray scale and depth images, and also different views of the scene. As such they can be used for training deep learning methods for in-car specific tasks like driver activity recognition[1][2][3][4], driver pose recognition[1], driver gaze detection[3][5], distracted driver recognition[2] and also for general vision tasks like object detection and semantic segmentation. An immediately apparent deficiency in all these datasets is that they always record only the driver in the scene and not the passenger. None provide recorded images for children/infants and child seats. This information is very useful for automobile manufacturers to adjust airbag deployment and to ensure child safety practices are being followed. The dataset recorded at [DFKI] fills these gaps in currently available in-cabin datasets by providing depth images from a wide angle front view. Apart from this factor, our dataset provides 3D single frame as opposed to activity annotations as is mostly the case with in-cabin datasets.

2 Deviation from objectives and corrective actions

The quantity of recorded data is smaller than originally planned due to Covid-19 restrictions. As a corrective action, the recorded data have been complemented with a comprehensive set of synthetic in-cabin data, including persons and child seats. The generation of these synthetic data was based on simulation setup developed in a joint project between [IEE] and [DFKI]. First simulations of rear bench scenarios had been already published earlier (see

[1] M. Martin et al., "Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles," 2019 IEEE/CVF International Conference on Computer Vision (ICCV) 2019
[2] M Selim et al. “AutoPOSE: Large-scale Automotive Driver Head Pose and Gaze Dataset with Deep Head Orientation Baseline.” VISIGRAPP (2020)
[3] A. Jain et al. “Brain4Cars: Car That Knows Before You Do via Sensory-Fusion Deep Learning Architecture”, arXiv:1601.00740 (2016).
[4] E. Ohn-Bar et al. “Head, eye, and hand patterns for driver activity recognition. In International Conference on Pattern Recognition (ICPR), 2014
[5] R. F. Ribeiro and P. Costa, “Driver Gaze Zone Dataset With Depth Data”, IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)

https://sviro.kl.dfki.de/data/). For the VIZTA data set described in this deliverable, the simulation tool had been developed further to also generate 3D data and has been adapted to the more complex in-cabin scenarios on the front seats corresponding to the real in-cabin test setup.

As a result, the total amount of data in this deliverable corresponds approximately to the initially planned data. The only remaining minor deviation is that the number of child seat recordings is still rather small and may be further increased during the next months.

3 Impact of the results

The data set described in this delivery is the first set of in-cabin recordings that covers the full in-cabin front scene from a single view of wide-angle time-of-flight camera. The 2D and 3D annotation of persons and objects allows the development of the core function of in-cabin scene understanding, which is the occupancy detection and segmentation.

As a basis for the training and evaluation of deep learning algorithms, this data set is a crucial milestone in the development of the AS2 demonstrator within VIZTA. Additionally, **[DFKI]** plans to make the dataset publicly available after the acceptance of a related publication. This action will encourage further researches in the topic and provide a benchmark for the evaluation of in-cabin detection and segmentation methods. For the VIZTA project, making the dataset openly available will also be an important public dissemination action. Finally, the use of synthetic and real data in the dataset will provide an excellent (and currently missing) opportunity for evaluation of domain adaptation methods to the scientific community.

4 Related IPR

Not applicable.

Deliverable report

1 Introduction

This report describes in detail a comprehensive in-cabin data set consisting both of recorded and simulated data. The use of the car-simulator setup located at [DFKI] for performing the recordings is presented, as well as the specific cases and parameter ranges covered in the dataset, naming conventions of the data and annotation. Annotation of the data was done at [DFKI] using a 3D annotation tool developed at [DFKI] that will allow faster labelling of large amounts of data.

Compared to the previous deliverable D3.33 the amount of real data has been expanded both in number of recordings and scene variations, comprising now also child seats and more objects. The synthetic data are entirely new, generated with a simulation setup adapted to the real in-cabin scenarios. The developed simulation environment is therefore also presented in detail in this report.

Machine Learning approaches using neural networks that are being currently developed at [DFKI] within the VIZTA project rely heavily on large amounts of representative data. Therefore, being able to use this data for training and evaluation of algorithms is an important step towards achieving the goals of the project.

The annotated dataset will be used to advance the research of [DFKI] and [IEE] in the in-cabin use-case scenarios of VIZTA. Equally importantly, the dataset will be published and made available to the scientific community for promoting further research and a benchmark for in-cabin estimation from ToF data. The dataset will cover an existing unfulfilled requirement in data from a single wide-angle ToF camera view covering both driver and passenger seats. Additionally, the co-existence of real and synthetic data in datasets is rarely encountered and provides excellent opportunities for testing domain adaptation approaches when neural networks are trained on synthetic data and tested with real data.

2 In-cabin data set

2.1. Recording setup and procedure

2.1.1. Data recording setup

For data recording, [DFKI] uses the in-cabin test platform documented in D3.31 (see Figure 1). Briefly explained, it consists of a realistic in-cabin mock-up, equipped with a wide-angle projection system for a realistic driving experience.

Figure 1: The DFKI in-cabin test setup

An Azure Kinect camera with a wide field of view is mounted at the rear mirror position for 2D and 3D data recording. The captured data consists of RGB, depth and IR amplitude images. To ensure a wide range of variability in the data set, the seat positioning of the simulator is adjustable via a CAN bus system.

2.1.2. Recording procedure and tools

[DFKI] has defined specific data acquisition rules alongside with a set of special recording tools to standardize the recording process.

Prior to each recording, the external calibration is evaluated with a specially designed checkerboard pattern that can be mounted in a fixed pre-defined position in the driving simulator setup (see the RGB image in Figure 2). Prior to each recording session, the position of the camera is verified, and in addition an extrinsic calibration performed. Therefore **[DFKI]** has developed a tool that calculates the rotation matrix and translation vector of the camera with respect to the checkerboard. These extrinsic calibration parameters are also exported via YAML file and provided to the annotation tool.

The recording engineer is supported by several in-house developed tools, helping with instructions to facilitate the application of the test matrix and with automatic filename generation for easily information tracking of test configurations and scenarios (see section 2.1.7).

After recording, **[DFKI]** has extracted the data from the raw format files and labelled them. Therefore, **[DFKI]** has developed an extraction tool that provide PNG data of each single image of recorded sequences; the details on the data format are described below in section 2.1.4). The associated filename is provided with all necessary information to identify its features (see section 2.1.3). For labelling of the data, a designated tool developed at **[DFKI]** is used (see section 2.1.5).

2.1.3. Recording definition

[DFKI] has developed a test matrix to guarantee the full traceability of the recorded data and to ensure that all variables are changed in a clearly defined way. These variables are:

- person or persons recorded
- accessories of persons
- choreography (dynamic and static poses)

- objects on seat (different instances per object and poses)
- seat position (consists of three independent variables)

For each of those variables **[DFKI]** has created lists of already used and planned instances. In the test matrix, these variables are mixed in a way that there is no correlation between any variables. Therefore, the seat position parameters are simultaneously varied in a quasi-random manner such that after any number of recordings their distribution corresponds to a random equal distribution. For objects the position and pose were randomly varied when repeating the recording in different seat positions. For the child seat recordings, **[IEE]** had provided representative samples of different categories. The seats were recorded in different configurations according to manufacturers' instructions and both empty as well as occupied by an infant or child doll of appropriate size.

[DFKI] and **[IEE]** agreed on a filename convention, which contains all necessary information about each recording configuration and scenario. This string comprises date and time, the matching calibration files, person or object ID, seat positioning code, accessory ID or object instance ID, choreography ID and an extra code block for use with later scenarios like body pose tracking.

2.1.4. Raw data formats

The format of the image data is the following.

- The video stream of the Azure Kinect is compressed in the Matroska container format MKV.
- For easier access to the data, the individual video frames are extracted into PNG images.
- Regarding the RGB data, the PNGs are three channels 8-bit images.
- The IR amplitude and depth images are stored as 16-bit encoded gray values. The depth value is stored in millimeters. Invalid depth measurements are coded with value 0.

For a better visual representation of both the amplitude and the depth images, **[DFKI]** developed tools to convert them. In case of the depth imagery, a false color representation is used where blue is close, and red is far away from the camera. Invalid or out of range pixels (value 0) are shown black. Regarding the amplitude images, the square root is calculated, and the image is equalized for better visualization.

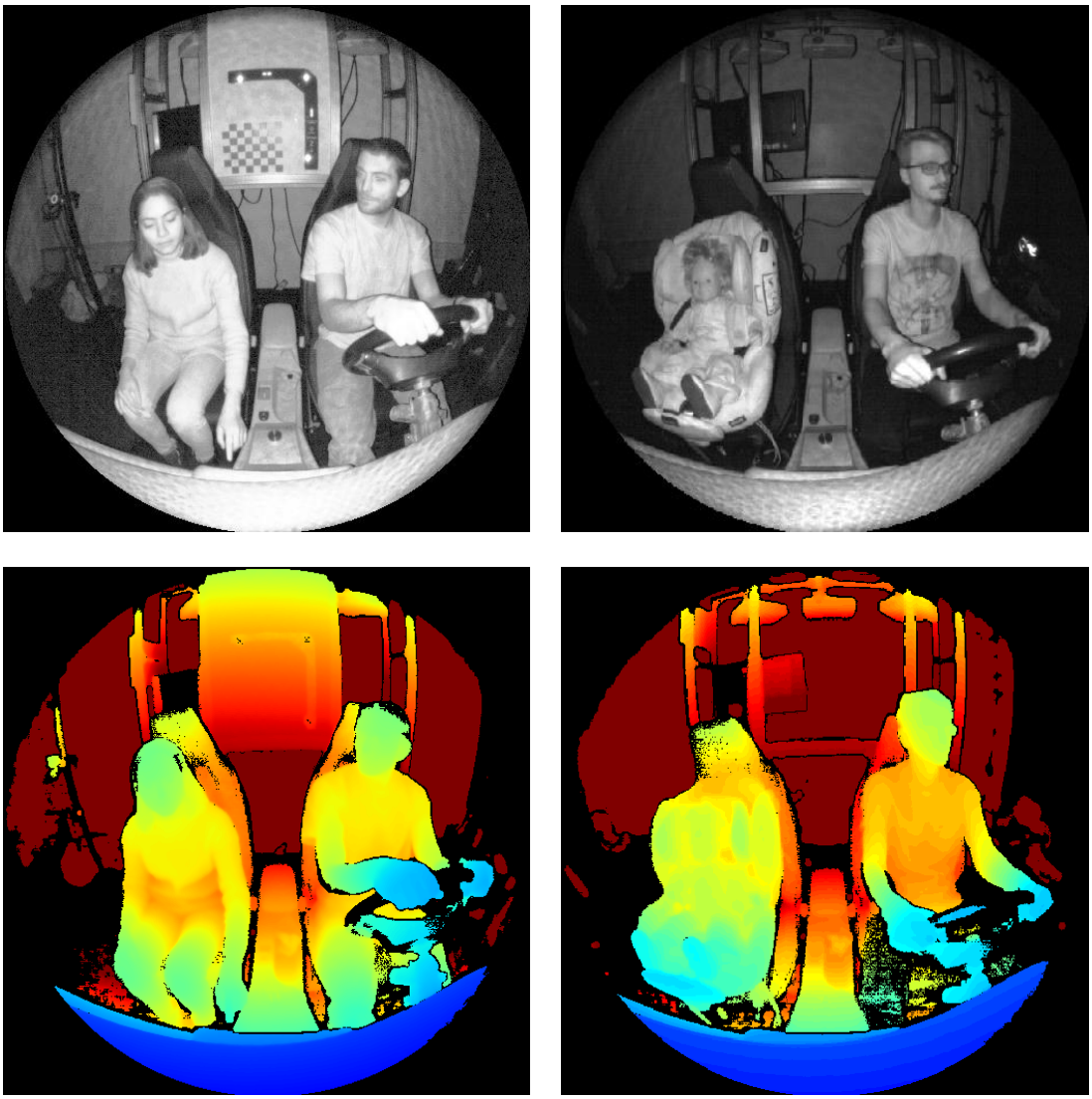

Figure 2: Visualization of the raw data extracted from two different recordings: Left Columnt: Person on passenger seat; Right Column: Front-facing child seat with child doll, Top Row: RGB-image. Middle Row: 16-bit depth image in false-color representation. Bottom row: 16-bit IR amplitude image, with some equalization for visualization

2.1.5. Annotation

Annotation of the data is done using the RGB-D annotation tool developed at **[DFKI]**. This tool allows a semi-automatic annotation mechanism of RGB-D image-sequences. The main functionalities the tool provides are (see Figure 3-5 generated by the tool):

- Un-distortion of RGB images and depth maps based on camera parameters. The resulting un-distorted depth map has a higher resolution than the original image and is sparse such that each valid pixel corresponds exactly to one 3D-point of the original distorted depth image (see Figure 4)
- 2D bounding box annotation of objects in RGB images and depth maps
- 3D bounding boxes annotations through an interactive 3D point cloud viewer
- Pixel segmentation of objects in RGB and depth images as well as 3D points segmentation

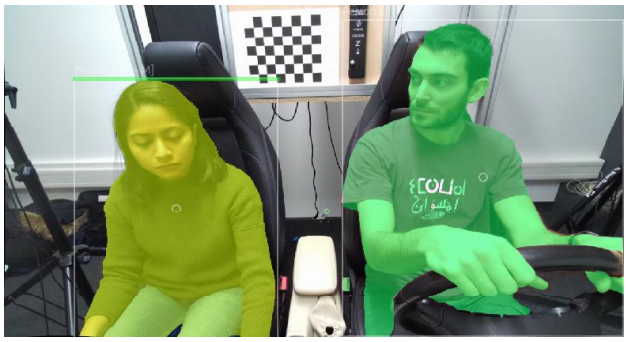

Figure 3: Undistorted RGB image overlaid with bounding boxes and instance segmentation mask

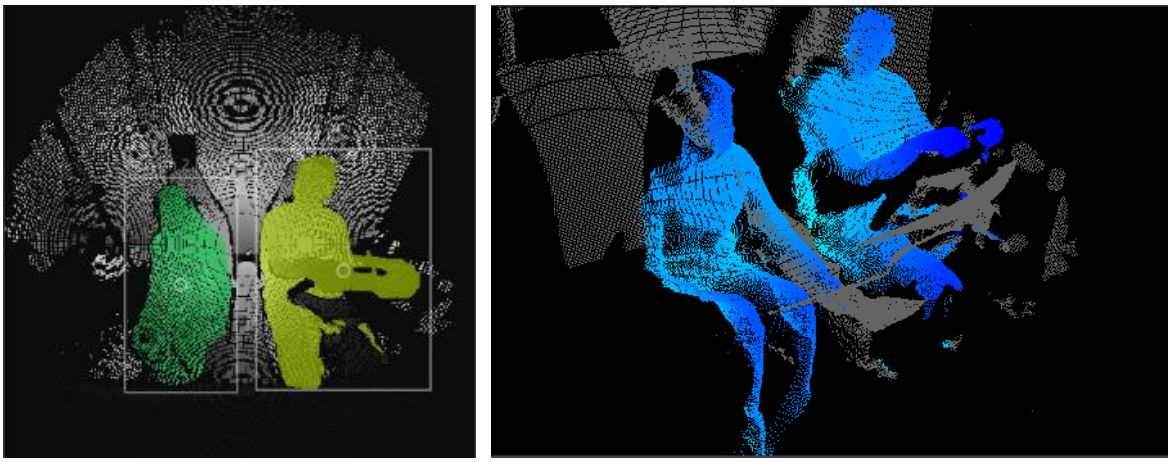

Figure 4: Left: Undistorted depth image with 2D bounding boxes and pixel segmentation. Depth values are shown as grey values. Segment colors correspond to instance ID. Right: Point cloud of segmented depth pixels. Color of 3D points corresponds to depth values

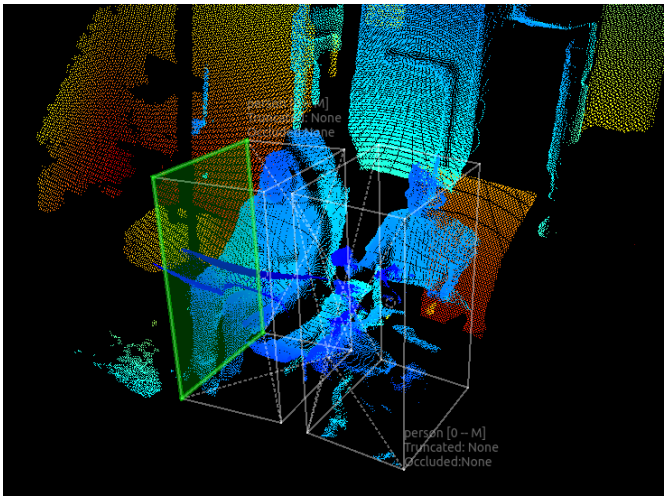

Figure 5: 3D bounding boxes around the point clouds of two persons in the cabin. The color map of the 3D point cloud is determined by the corresponding depth values. Each box has as attributed a class label

The format of the annotation data, as provided in the data set, is described in section 3.

2.1.6. Post-processing of recorded data

The post-processing of the data consists of a rectification and a normalization. In this way, the recorded data provided in the data set are consistent with the simulated data.

Image Rectification: All recorded depth images are undistorted using the OpenCV[6] functions *cv2.remap* and *cv2.initUndistortRectifyMap* function to remove lens distortion. Therefore, a pinhole camera model field of view of 106°x106° with 512 x 512 pixels has been defined to which all images are mapped. Additionally, since the camera orientation is not exactly the same for all the images in the dataset, all the images are mapped to a common camera pose by considering the extrinsic calibration parameters in the image rectification. The resulting mapping has been applied both to the extracted depth and amplitude images, as well as to the annotations, i.e. the segmentation masks and 2D boxes. An example of this rectification process is shown in Figure 6. The 3D boxes need not to be mapped, as they refer to the world coordinate system. The RGB-image is also undistorted using the intrinsic provided by the camera manufacturer. The resolution and field of view is, however, different to that of the depth camera. (see section 3.1).

Normalization: The raw depth images contain depth values up to a range of 3.5m which is not entirely required. Also there are pixels which have erroneous very high values of depth. Therefore, we clip all depth values to the limit of 2550 millimeters such that all values lie within the range [0,2550]. We choose this number as the rough limit of the depth of interest in our driving simulator and also to ease image conversions. We finally save all images in 16bit format.

[6] Open Source Computer Vision Library https:\\opencv.org

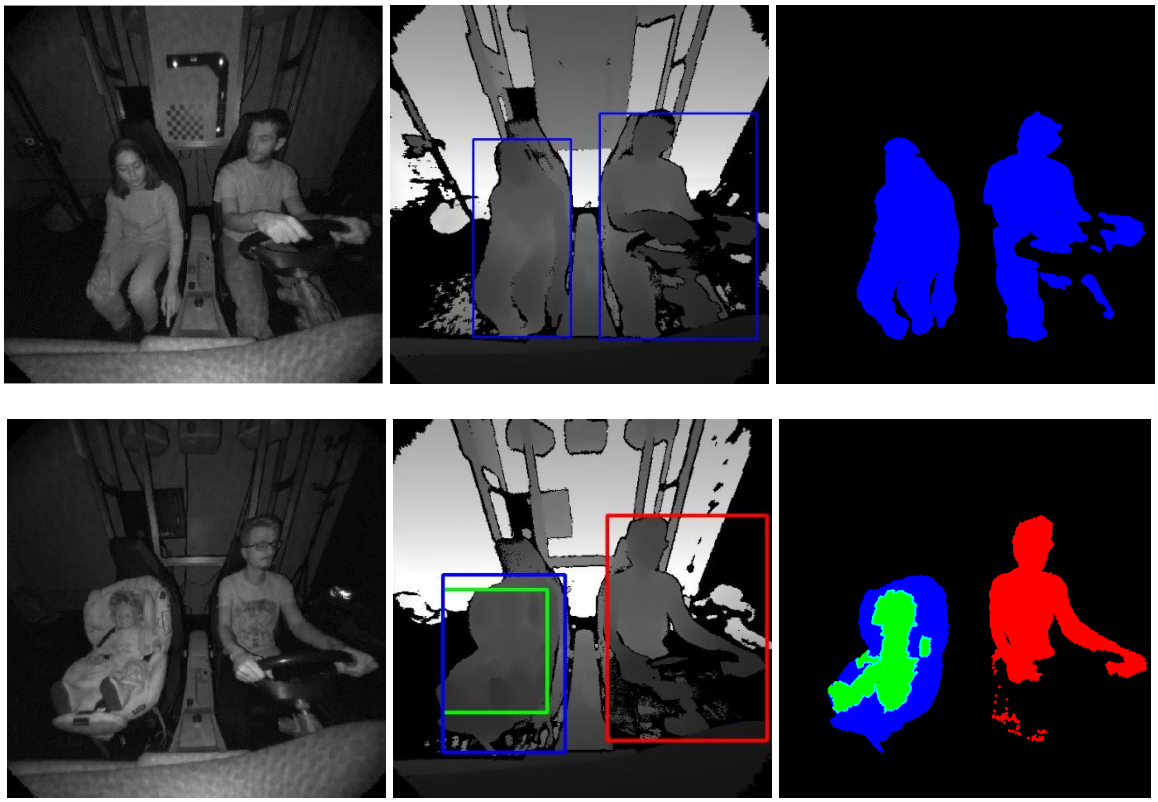

Figure 6: Depth camera output and annotation after post-processing for two different scenes shown in Figure 2 Left column: Undistorted IR amplitude image, Middle column: Undistorted normalized depth image in gray scale representation overlayed with the mapped 2D boxes, Right column: Undistorted pixel class masks

2.1.7. Recorded Scenarios

As described in section 2.1.3 the recorded scenarios have been defined by a set of pre-defined variables, which are documented in the filename to guarantee full traceability. Most important variable is thereby the type of occupancy, being person or object as well as the pose and action of persons.

For the person recordings, some choreographies have been defined consisting of several typical actions of poses a driver performs while driving. These are:

- Drive normal, Grasping the steering wheel with both hands
- Looking to the left (with turning steering wheel left)
- Looking to the right (with turning steering wheel right)
- Operate navigation screen / change gears
- Reaching for the glove compartment
- Moving hand to head
- Leaning forward
- Turning upper torso left
- Turning upper torso right
- Reversing the car (turning upper torso extremely right)

In addition, there are choreographies defined for passengers which are:

- Sitting normal
- Handling navigation screen

- Looking out of window
- Handling sun visor
- Turning backwards,
- Turning upper torso left
- Turning upper torso right
- Talking to driver
- Taking something from dashboard
- Reading paper/book

2.1.8. Recorded data statistics

Table 1 gives an overview of the recorded and annotated data. It should be noted that recordings with persons consist of sequences of typically 30sec or more, containing a whole choreography of typical person movements and gestures while driving (see section 2.1.7). By contrast, object and empty scene recordings lasted only 10 frames (less than 1sec). To keep the annotation effort limited, only every 20th frame was annotated which corresponds to 1.5 annotated frames per second of recording.

Table 1: Distribution of sequences and annotated frames over the different front seat occupancy scenarios

Occupancy Scenario				
Driver Seat	Passenger Seat	Number of sequences	Average Sequence Length [sec]	Number of annotated Frames
Person	Empty	48	31	2327
Person	Object	28	30	1303
Person	Rear-facing Infant Seat	11	43	704
Person	Front-facing Child Seat	16	43	1054
Person	Person	12	50	924
Object	Empty	178	1	356
Empty	Empty	240	1	0
Total		**533**		**6668**

The following Tables Table 2 - Table 4 give an overview over the number of different persons, objects and child seats used in the recordings summarized in Table 1.

Table 2: Distribution of ethnics and gender of the 13 test persons

	Male	Female	Total
Asian	2	2	4
Caucasian	7	2	9
Total	9	4	**13**

Table 3: List of different objects categories as well as number of different objects instances and recordings per category. The Object IDs Oo0 and Oo1 correspond to empty seats and therefore not listed here.

Object Category	Object ID	Number of different object instances	Number or sequences
Backpack small	Oo2	3	25
Winter jacket	Oo3	4	22
Box small	Oo4	3	24
Water bottle	Oo5	4	13
Mobile phone	Oo6	3	10
Blanket	Oo7	2	20
Small cloths (wooly hat, scarf, gloves)	Oo8	5	19
Book	Oo9	5	22
Laptop	O10	3	9
Laptop bag	O11	2	8
Backpack large	O12	3	13
Handbag	O13	5	21
Total number	**12**	**42**	**206**

Table 4: List of different child seat categories as well as number of different seats and configurations per category and number of recordings. Sequences were recorded with child seats empty or occupied with an infant/child doll.

Child seat category	Number of different seats	Number of different seat configurations	Number of sequences
Rear-facing infant seat	2	RF RF+handle up RF+sunshield	3 3 1
Convertible child seat	2	RF FF	4 4
Front-facing child seat	2	FF	12
Total	**6**		**27**

Table 5: Distribution of annotated subjects over different occupancy classes

Occupancy Class	Number or annotated subjects
Person	7236
Object	1658
Rear-facing infant seat (RF)	704
Forward facing child seat (FF)	1054
Infant (in RF)	379
Child (in FF)	448
Total	**11497**

In total 485 sequences have been recorded with a total length of 71 minutes. From the 128.000 frames, 6456 have been annotated, resulting in 11497 annotated instances (see Table 5). The effort to annotate the frames was in total approximately 590 person hours.

2.2.Simulation Setup and procedure

We generate a synthetic dataset for vehicle front-seat scenarios to be used for training generalizable models for car in-cabin scene understanding. Blender[7], an open-source 3D rendering tool, was used for simulating real-life scenarios that could be difficult to simulate in real data recordings. Additionally, we used SVIRO[8] a synthetic dataset for the passenger rear seat compartment in different vehicles, as our base project. This dataset has been the result of a joint project between VIZTA partners [**DFKI**] and [**IEE**] and has been modified by [**DFKI**] to capture instead the front-seating of cars, taking into consideration the different poses, items, layout and camera pose and orientation that differ between the front and back of the vehicle.

2.2.1. Simulation Tools and Setup

Blender[7] is an open source 3D creation suite that supports the entirety of the 3D pipeline from modeling, simulation and composition. We have used Blender to embed the models and synthetic objects obtained from the sources mentioned in this section into our scene, and used its python API to have a full control over the coordinates and orientation of the different objects as well as randomizing their different combinations.
MakeHuman[9] is an open source 3D computer graphics software designed to simulate realistic human models. We used it to generate a number of human models along with their clothing in a random fashion

[7] Blender 3d rendering tool. https://www.blender.org
[8] Steve D. Da Cruz et al. “Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark, https://sviro.kl.dfki.de (2020)
[9] MakeHuman 3d rendering tool. http://www.makehumancommunity.org

Synthetic Objects. Different data sources were used to acquire the different components that make up a realistic car scenery. The 3D models of the cars were purchased from Hum3D[10], the everyday objects were downloaded from Sketchfab[11] and the human models were generated via MakeHuman as mentioned previously. In addition, High Dynamic Range Images (HDRI)[12] was used to get different environmental backgrounds and lightings, and finally in order to define the reflection properties and colors for the 3D objects, textures from Textures.com[13] were obtained for each object.

Car and Camera Model: Among the difference car model that were integrated in the simulation in the SVIRO[8] project, the Mercedes A-class was chosen here, as the in-cabin setup at **[DFKI]** has real front seats from this car model integrated. But please note that except the front seats, the interior of the in-cabin mock-up at **[DFKI]** does not correspond to the A-class model

Camera Model The camera intrinsic parameters in Blender were chosen according to the pinhole camera model used in the rectification of the real data (see section 2.1.6). The position and pose of the virtual camera were adjusted such that the field of view corresponds approximately to that in the real in-cabin scene at **[DFKI]**. To simulate the perception of the scene by a camera with active illumination a light source located at the cameras position was add to the simulation model. The 3D rendering provided thus photorealistic intensity images taking the objects reflection properties (albedo, roughness) and geometry (depth, normal) into account. The material properties in the infrared are, however not yet included in the rendering software yet. Therefore, the red channel was used instead to "imitate" the infrared image. For more details, see[8]. The photon shot noise is not included in the simulation and also not time-of-flight specific artefacts in the depth measurement, as e.g., motion artefacts.

2.2.2. Simulation output

A synthetic dataset with 3200 static sceneries for the Mercedes-Benz A-Class vehicle is generated with each scenery containing following data and annotations.

- Depth map (PNG format) and raw depth map (EXR format)
- RGB Image (PNG format)
- Combined Instance and class segmentation masks (PNG format)
- Bounding boxes (TXT format)
- Key points for pose estimation (JSON format)
- Imitated IR greyscale image (PNG format): This grey image does not exactly correspond to a ToF-camera infrared amplitude image, as material properties in the infrared are not included in rendering software yet. For more details, see[8]

[10] Hum3d 3d rendering tool. http://www.hum3d.com
[11] Sketchfab 3d rendering tool. http://www.sketchfab.com
[12] HDRIHaven 3d rendering tool. http://www.hdrihaven.com
[13] Textures.com 3d rendering tool. http://www.textures.com

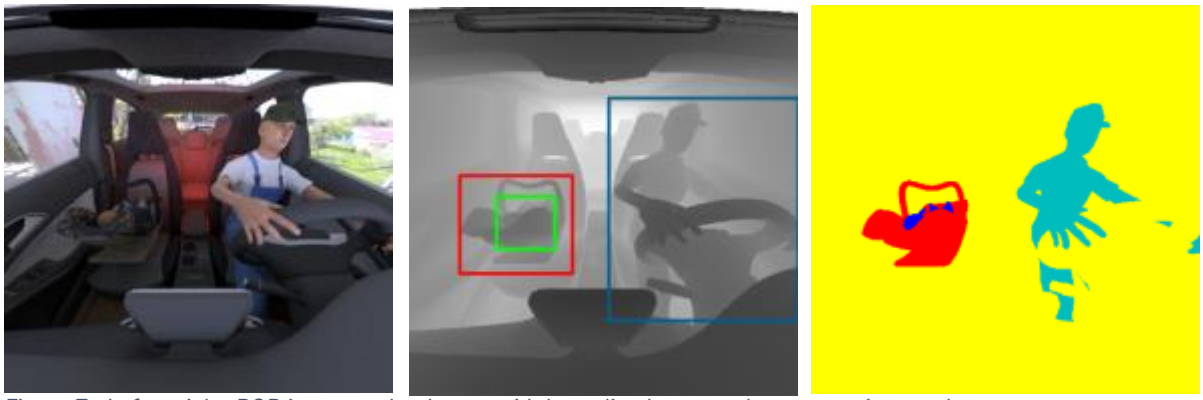

Figure 7:: Left to right: RGB image, a depth map with bounding boxes and segmentation mask

2.2.3. Post-processing of simulated data

The simulated data did not require any post-processing but only minor format changes to make them consistent with the real data (after postprocessing). The depth images in float format have been normalized to 1mm resolution and converted to 16bit integer values as the depth values from the Kinect. The annotations have been converted to the same format as for the real data. See the description of the data formats in section 3.

2.2.4. Simulated Scenarios

In order to replicate real driver and passenger poses, the joints and limbs of human models in Blender are manipulated to look similar to the human poses in recorded real dataset. Since the driver poses are always restricted by the car elements they are interacting with, and in order to avoid intersection of human and car models in Blender, some fixed poses are created for the hand positions, that are possible in real driving scenarios. Following poses are simulated across the different driver models in our dataset:

- Looking to their right and left
- Grasping the steering wheel with both hands
- Steering to their right and to their left
- Leaning forward
- Handling the music player/navigator
- Reaching for the glove compartment
- Reversing the car
- Waving
- Being on the phone
- Reaching for an object on the passenger seat
- Giving a right/left signal

On the other hand, there were no such restrictions for the passenger poses, and thus we replicated the poses used in SVIRO by randomly selected body poses within the constraint of the seating.

2.2.5. Simulated data statistics

The total 3306 sceneries have been generated. In all cases the driver seat was occupied by a person, while for the passenger seat, the occupancy was generated at random. Table 6 shows the distribution of the number of instances for each class on the passenger seat. The total number of annotated subjects on driver or passenger seat sums to 6974. The number of people and distribution of the gender, age and ethnicity for the data set can be found in Table 7.

Table 6: Distribution of occupancy classes on the passenger seat. Note that the number of infant and child is also included in the number of rear-facing infant seat and forward facing child seat

Passenger Seat Occupancy	Number of Instances
Infant (in rear-facing infant seat)	439
Child (in forward-facing child seat)	380
Adult	806
Object	338
Rear-facing infant seat	878
Front-facing child seat	827
Empty	457
Total	**3668**

Table 7: Distribution of ethnicities and gender between the different human models.

	Adult	Child	Baby
African	2	2	1
Asian	2	2	1
Caucasian	2	2	1
Male	3	3	-
Female	3	3	-
Total	6	6	3

Summarizing the simulated data statistics with the statistics of recorded data in 2.1.8 one finds that the data set comprises in total 9974 sceneries (6668 recorded plus 3306 simulated) with 18471 annotated subjects (11497 in recorded frames and 6974 in simulated frames). The splitting of the data into training and testing data will be provided when the data set is made public available for benchmarking.

3 Data format

Both the real and synthetic data are delivered in the same format for both image data and annotations with a few differences explained below.

3.1. Image data

The camera data provided are in detail:

- **Depth Z-Image**. The depth image is undistorted with a pixel resolution of 512 x 512 pixels. The depth values are normalized to [1mm] resolution and clipped to a range of [0,2550mm]. Invalid depth values are coded as '0'. Images are stored in 16bit PNG-format.
- **IR Amplitude Image**. Has the same format as the depth image.
- **RGB Image:** Undistorted color images are saved in PNG-format in 24bit depth. While the simulated RGB images have the same resolution as the corresponding depth images, the recorded RGB images have a higher resolution of 1280x720 pixels, but a lower field of view of 90°x59° FOV (see also the description of the in-cabin test platform D3.31).

The dataset of this deliverable comprises annotation of the 3D data. The annotation formats are:

- **2D bounding boxes** are defined by the (x,y) image coordinates of the top-left corner, the (w,h) width and height of the box, and its class label which is either "person", "child", "infant", "RF", "FF" or "object". In case of a child in a child seat, both the child and the child seat are annotated. For the class "object" there are 12 different object categories defined according to Table 3, which are also encoded in the filename of the recording together with all other relevant parameters of the recording and test setup. The filename encoding is provided in a separate document with the data.
- **3D bounding boxes**: Each 3D box is represented by the coordinates (cx, cy, cz) of the box center, its dimensions (width, height, depth), its orientation along x-, y- and z- axes with respect to the world coordinate system. In addition, each box is annotated with its class label. In addition, there are flags to denote if the box is occluded or truncated.
- **Pixel segmentation masks:** For 2D pixel-level segmentation, each segmented object has an associated class and instance ID. For each depth image two corresponding masks are generated: instance mask and class mask. In each of these masks, the pixel intensity corresponds to the pixel or class ID. As there is a one-to-one correspondence between depth image and point cloud, the image mask provides also point cloud segmentation.

The annotations are stored in CSV format, except the segmentation masks which are stored as PNG-images.

4 Publishable information

[DFKI] has published the extensive labelled dataset of vehicle in-cabin scenarios on the website https://vizta-tof.kl.dfki.de/ hosted at their servers. The intention of the open access to these data is to foster the research on in-cabin monitoring function. A conference publication presenting this dataset is in preparation; a first version has been uploaded to the e-print server arXiv.org under the ID arXiv:2103.11719 as reference to the dataset.

5 Conclusion

In this deliverable report, a complete dataset of in-cabin scenarios is described which consists in both recorded and synthetic data. The data are used for the development and testing of the core algorithms corresponding to in-cabin monitoring scenario of VIZTA. It also allows benchmarking of different methods in order to select the most promising for the in-cabin VIZTA final demonstrator. The procedure of capturing data covering different occupancy scenarios was also described - as well as the labelling of the data using a [DFKI] 3D Annotation Tool. Additionally, the dataset is complemented by synthetic data generated with a simulation environment that had been jointly developed by [DFKI] and [IEE]. The generation of synthetic data increased both the size and value of the dataset. It covers now approximately 10.000 annotated frames (6668 from recorded sequences plus 3306 simulated static scenarios) with approximately 18.500 annotated subjects (11497 in recorded frames and 6974 in simulated frames) The value is also increased because the dataset can not only serve as benchmark for in-cabin algorithm development but also serve as a basis for the development of domain adaptation and transfer learning methods. In the future the dataset may be further expanded by simulating various car interiors and adding artefacts as background light influence on both simulated and reals data.