Driver and Pedestrian Mutual Awareness for Path Prediction in Intelligent Vehicles

Roth, M.

DOI
10.4233/uuid:7d3c6107-812e-458e-bd11-04f5c1e5931a

Publication date
2023

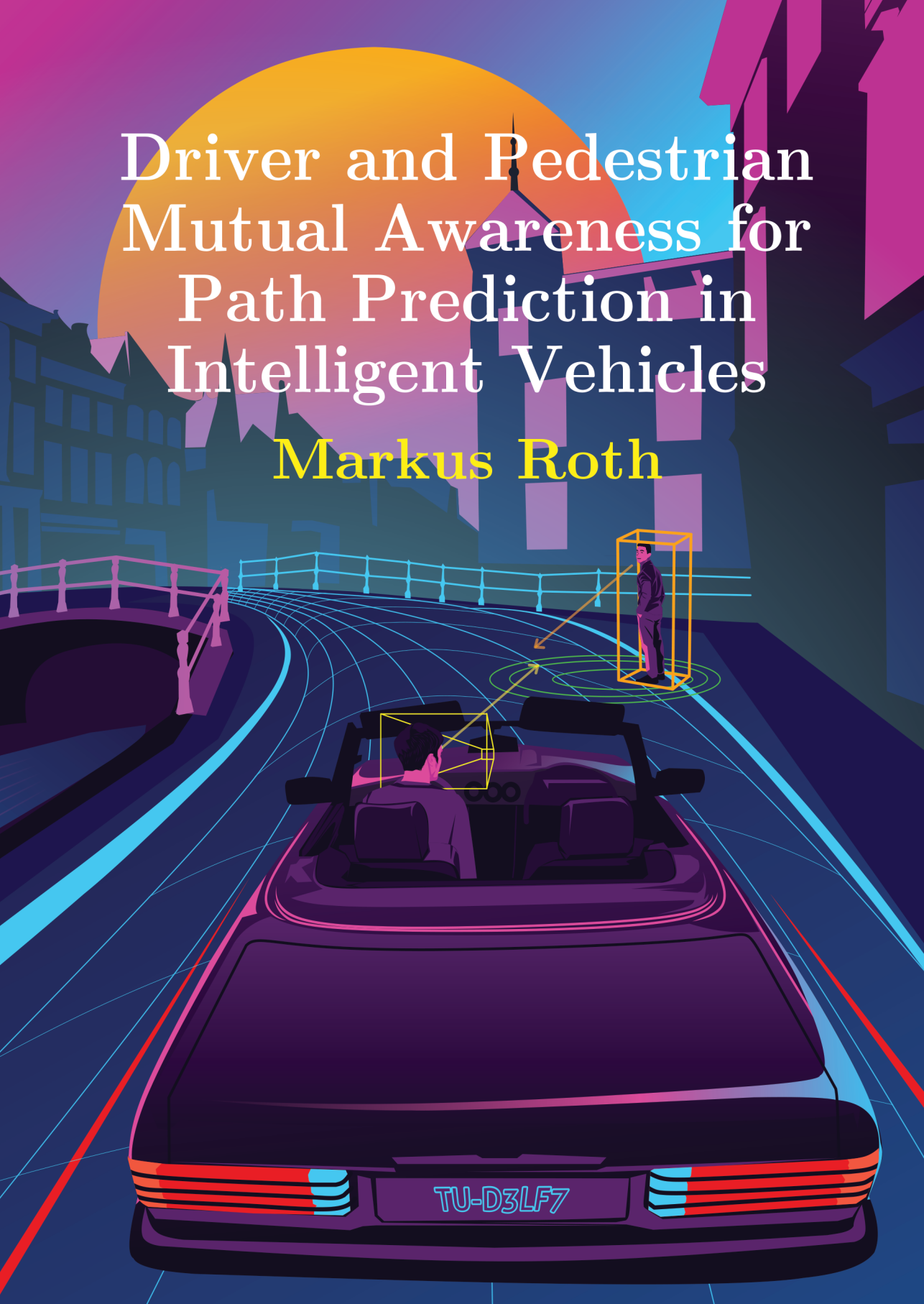
Document Version
Final published version

Citation (APA)
Roth, M. (2023). *Driver and Pedestrian Mutual Awareness for Path Prediction in Intelligent Vehicles*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:7d3c6107-812e-458e-bd11-04f5c1e5931a

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Driver and Pedestrian Mutual Awareness for Path Prediction in Intelligent Vehicles

Markus Roth



Driver and Pedestrian Mutual Awareness for Path Prediction in Intelligent Vehicles

DRIVER AND PEDESTRIAN MUTUAL AWARENESS FOR PATH PREDICTION IN INTELLIGENT VEHICLES

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates,
to be defended publicly on
Wednesday, 20 December 2023 at 17:30 o'clock

by

Markus ROTH

Diplom-Informatiker,
Karlsruhe Institute of Technology, Karlsruhe, Germany,
born in Talmesch, Romania.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	Chairperson
Prof. dr. D.M. Gavrila	Delft University of Technology, promotor
Dr. J.F.P. Kooij	Delft University of Technology, copromotor

Independent members:	
Prof. dr. G.C.H.E. de Croon	Delft University of Technology
Prof. dr. ir. J.C.F. de Winter	Delft University of Technology
Prof. dr. C. Wöhler	TU Dortmund University, Germany
Prof. dr. J.M. Zöllner	Karlsruhe Institute of Technology, Germany
Dr. H. Caesar	Delft University of Technology

Keywords: Head pose estimation, Head pose dataset, Person detection, Ego-vehicle path prediction, Pedestrian path prediction, Intelligent vehicles, Automated driving

Cover art: Markus Roth (concept), Ridzwan Irawan (illustration)

ISBN 978-94-6384-502-1

An electronic version of this dissertation is available at
https://repository.tudelft.nl/.

You can always be good,
but experience comes with time.

Reinhold Roth

CONTENTS

SUMMARY

This thesis addresses the sensor-based perception of driver and pedestrian to improve joint path prediction of ego-vehicle and pedestrian based on mutual awareness in the domain of intelligent vehicles.

According to the World Health Organization (WHO), more than half of global traffic deaths are among Vulnerable Road Users (VRUs), such as pedestrians and riders, and human error is still a major cause of accidents. This motivates paying special attention to pedestrians and drivers while they are interacting in traffic. For the foreseeable future, the reality on the road (and the accident numbers) will largely be determined by Advanced Driver-assistance Systems (ADAS) where the driver is still required to keep the eyes on the road. To that end, the scope of this thesis resides within ADAS and driving automation up to (including) autonomy level 3 as defined by the Society of Automotive Engineers (SAE). While current ADAS consider pedestrians and the driver individually, their mutual awareness has not been leveraged to improve path prediction and thereby road safety.

This thesis presents a framework that estimates **driver head pose** from driver camera images, estimates **pedestrian location and orientation** from exterior camera images and lidar point clouds, uses this information over time to reason about **driver and pedestrian mutual awareness**, and performs **joint probabilistic path prediction of ego-vehicle and pedestrian** to assess **collision risk**.

Deep neural networks demand a large training set to tune the vast amount of parameters. This thesis introduces DD-Pose, the Daimler TU Delft Driver Head Pose Benchmark, a large-scale and diverse benchmark for image-based head pose estimation and driver analysis. It contains 330k measurements from multiple cameras acquired by an in-car setup during naturalistic drives. Large out-of-plane head rotations and occlusions are induced by complex driving scenarios. Precise head pose annotations are obtained by a motion capture sensor and a novel calibration device. The new dataset offers a broad distribution of head poses, comprising an order of magnitude more samples of rare poses than a comparable dataset.

Utilizing the dataset, this thesis presents intrApose, a novel method for continuous 6 degrees of freedom (DOF) head pose estimation from a single camera image without prior detection or landmark localization. intrApose uses camera intrinsics consistently within the deep neural network and is crop-aware and scale-aware: poses estimated from bounding boxes within the overall image are converted to a consistent pose within the camera frame. It employs a continuous, differentiable rotation representation that simplifies the overall architecture compared to existing methods. Experiments show that leveraging camera intrinsics and a continuous rotation representation ($SVDO^+$) results in improved pose estimation compared to intrinsics agnostic variants and variants with discontinuous rotation representations. Driver head pose of naturalistic driving is biased towards close-to-frontal orientations. Training with an unbiased data distribution, i.e., a

more uniform distribution of head poses, further reduces rotation error, specifically for extreme orientations and occlusions.

In addition to considering the inside of the vehicle, this thesis also focuses on the outside environment and presents a method for 3D person detection from a pair of camera image and lidar point cloud in automotive scenes. The method comprises a deep neural network that estimates the 3D location, spatial extent, and yaw orientation of persons present in the scene. 3D anchor proposals are refined in two stages: a region proposal network and a subsequent detection network. For both input modalities high-level feature representations are learned from raw sensor data instead of being manually designed. To that end, the method uses Voxel Feature Encoders to obtain point cloud features instead of widely used projection-based point cloud representations. Experiments are conducted on the KITTI 3D object detection benchmark, a commonly used dataset in the automotive domain.

Eventually, the output provided by the methods of the former chapters, namely, driver head pose and 3D person locations, are leveraged by a novel method for vehicle-pedestrian path prediction that takes into account the awareness of the driver and the pedestrian of each other's presence. The method jointly models the paths of ego-vehicle and a pedestrian within a single Dynamic Bayesian Network (DBN). In this DBN, sub-graphs model the environment and entity-specific context cues of the vehicle and pedestrian (incl. awareness), which affect their future motion. These sub-graphs share a latent state which models whether the vehicle and pedestrian are on collision course. The method is validated with real-world data obtained by on-board vehicle sensing, spanning various awareness conditions and dynamic characteristics of the participants. Results show that at a prediction horizon of 1.5 s, context-aware models outperform context-agnostic models in path prediction for scenarios with a dynamics change while performing similarly otherwise. Results further indicate that driver attention-aware models improve collision risk estimation compared to driver-agnostic models. This illustrates that driver contextual cues can support a more anticipatory collision warning and vehicle control strategy.

The main conclusions and findings of this thesis are: using a measurement device with a per-subject calibration procedure simplifies the data acquisition process to obtain a broad distribution of head poses. Using an intrinsics-aware head pose estimation method with a continuous rotation representations allows for a simple architecture that yields robust head pose estimates across a broad spectrum of head poses. Modeling of both driver and pedestrian mutual awareness in a unified DBN improves joint probabilistic path prediction compared to driver-agnostic models. Additionally, it provides explainability for model parameters and interpretability of the internal decision making process.

Further research can be conducted to understand the behavior of humans inside and outside an intelligent vehicle. Two major trends go towards integrating uncertainties into the components and combining them to a system that can be trained end-to-end from raw sensor data to predicted paths. Future work would greatly benefit from representative, worldwide, naturalistic, multi-sensor, temporal data which cover the outside environment as well as the inside of the vehicle – ideally shared across research institutions and companies.

SAMENVATTING

Deze dissertatie richt zich op het met sensoren waarnemen van bestuurders en voetgangers om de paden van het ego-voertuig en voetgangers gezamenlijk te voorspellen met in acht neming van wederzijdse gewaarwording, in het domein van intelligente voertuigen.

Volgens de Wereldgezondheidsorganisatie (WHO) valt meer dan de helft van alle verkeersdoden wereldwijd onder kwetsbare weggebruikers, zoals voetgangers en fietsers. Menselijke fout is nog steeds een belangrijke oorzaak van deze ongevallen. Dit motiveert om extra aandacht te besteden aan voetgangers en bestuurders tijdens hun interactie in het verkeer. In de nabije toekomst zal de realiteit op de weg (en de ongevallencijfers) grotendeels bepaald worden door geavanceerde bestuurders assistentie systemen (ADAS) waarbij de bestuurder nog steeds de ogen op de weg moet houden. Daarom ligt de focus van dit proefschrift op ADAS en automatisering van het rijden tot en met autonomieniveau 3 zoals gedefinieerd door de Society of Automotive Engineers (SAE). Hoewel de huidige ADAS rekening houden met voetgangers en de bestuurder afzonderlijk, is hun wederzijdse gewaarwording niet benut om de pad voorspelling en daarmee de verkeersveiligheid te verbeteren.

Deze dissertatie presenteert een systeem dat de **houding van het hoofd van de bestuurder** herleid uit camerabeelden, en de **locatie en oriëntatie van voetgangers** observeert uit externe camerabeelden en lidar puntenwolken. Over tijd wordt deze informatie gebruikt om door middel van een **gezamenlijke, probabilistische pad voorspelling van het ego-voertuig en de voetganger** te redeneren over de **wederzijdse gewaarwording tussen bestuurder en voetganger**, om het **risico op een botsing** in te schatten.

Diepe neurale netwerken vereisen een grote trainingsset om de enorme hoeveelheid parameters af te stemmen. Deze dissertatie introduceert DD-Pose, de Daimler TU Delft Driver Head Pose Benchmark; een grootschalige en diverse dataset voor beeld-gebaseerde observatie van hoofd bewegingen en bestuurder analyse. De benchmark bevat 330.000 metingen van meerdere camera's die zijn verkregen door een in-voertuig opstelling tijdens naturalistische ritten. Grote hoofdoriëntaties en occlusies worden veroorzaakt door complexe rijscenario's. Nauwkeurige annotaties van de hoofdhouding zijn verkregen door een bewegingsopnamesensor en een nieuw kalibratieapparaat. De nieuwe dataset biedt een brede distributie van hoofdhoudingen, met een orde van grootte meer zeldzame poses dan een vergelijkbare dataset.

Gebruikmakend van de dataset presenteert dit proefschrift intrApose, een nieuwe methode voor continue schatting van de hoofdhoudingen in 6 vrijheidsgraden uit een enkel camerabeeld zonder voorafgaande detectie of lokalisatie van herkenningspunten. intrApose neemt de intrinsieke camera-eigenschappen in acht binnen het diepe neurale netwerk en houdt rekening met het bijsnijden en schalen van afbeeldingen: oriëntaties geobserveerd in bounding boxes binnen het totale beeld worden geconverteerd naar een consistente oriëntatie in het cameraframe. De methode gebruikt een continue, differentieerbare rotatie representatie die de algehele architectuur vereenvoudigt

in vergelijking met bestaande methoden. Experimenten tonen aan dat het gebruik van camera-intrinsieke eigenschappen en een continue rotatie-representatie (SVDO+) resulteert in een betere oriëntatie bepaling ten opzichte van intrinsiek-agnostische technieken en varianten met discontinue rotatie representaties. Tijdens naturalistisch rijden is het hoofd van de bestuurder voornamelijk naar voren gericht. Trainen met een meer uniforme verdeling van hoofdoriëntaties vermindert bias-gerelateerde meetfouten, specifiek voor extreme oriëntaties en occlusies.

Naast observatie binnen het voertuig, richt dit proefschrift zich ook op de buitenomgeving en presenteert een methode voor 3D-persoonsdetectie uit een combinatie van camera beeld en lidar puntenwolk in verkeersscènes. De methode bestaat uit een diep neuraal netwerk dat de 3D-locatie, ruimtelijke omvang en oriëntatie van personen in de scène detecteert. 3D ankervoorstellen worden in twee fasen verfijnd: een regio voorstel netwerk en een daaropvolgend detectienetwerk. Voor beide invoer modaliteiten worden hoogwaardige representaties geleerd van ruwe sensor gegevens in plaats van handmatig ontworpen. Daartoe gebruikt de methode Voxel Feature Encoders in plaats van de veelgebruikte projectie-gebaseerde puntenwolken. Experimenten worden uitgevoerd op de KITTI 3D object detectie benchmark, een veelgebruikte dataset in de auto-industrie.

Vervolgens worden de resultaten uit de eerder benoemde methoden, namelijk, de houding van het hoofd van de bestuurder en de 3D-locaties van personen, gebruikt voor het voorspellen van het pad van voertuig en voetganger met een inachtneming van wederzijdse gewaarwording van elkaars aanwezigheid. De methode modelleert gezamenlijk de paden van het ego-voertuig en een voetganger binnen een Dynamisch Bayesiaans Netwerk (DBN). In dit DBN modelleren subgrafen de omgevings- en entiteit-specifieke contextuele informatie van het voertuig en de voetganger (inclusief gewaarwording), die hun toekomstige beweging beïnvloeden. Deze subgrafen delen een latente toestand die modelleert of het voertuig en de voetganger op ramkoers liggen. De methode is gevalideerd met praktijkgegevens die zijn verkregen door voertuig instrumentatie, die verscheidene gewaarwordingsomstandigheden en manoeuvres van de deelnemers omvatten. De resultaten laten zien dat bij een voorspellingshorizon van 1,5 s, contextbewuste modellen beter presteren dan context-agnostische modellen in pad voorspelling voor scenario's met een dynamische gedragsveranderingen, en vergelijkbaar presteren in andere gevallen. De resultaten geven verder aan dat gewaarwordingsbewuste modellen het risico op een botsing beter inschatten dan gewaarwording-agnostische modellen. Dit illustreert dat contextuele aanwijzingen van de bestuurder een anticiperende botswaarschuwing en voertuigcontrole kunnen ondersteunen.

De belangrijkste conclusies en bevindingen van dit proefschrift zijn: het gebruik van een meetapparaat met een kalibratieprocedure per participant vereenvoudigt het proces van gegevensverzameling van een brede verdeling van hoofdhoudingen. Het gebruik van een intrinsiek-bewuste methode voor het schatten van hoofdoriëntatie met een continue rotatie weergave maakt een eenvoudige architectuur mogelijk die robuuste meeting oplevert over een breed spectrum van houdingen. Modellering van wederzijdse gewaarwording tussen bestuurder en voetganger in een verenigd DBN verbetert gezamenlijke probabilistische pad voorspelling in vergelijking met bestuurders-agnostische modellen. Bovendien biedt het verklaarbaarheid voor modelparameters en interpreteerbaarheid van het interne besluitvormingsproces.

Verder onderzoek kan worden uitgevoerd om het gedrag van mensen binnen en buiten een intelligent voertuig. Twee belangrijke trends gaan in de richting van het integreren van onzekerheden in de componenten en ze combineren tot een systeem dat eind-tot-eind getraind kan worden van ruwe sensor gegevens tot voorspelde paden. Toekomstig werk zou veel baat hebben bij representatieve, wereldwijde, naturalistische, multi-sensor, temporele gegevens die zowel de buitenomgeving als de binnenkant van het voertuig bestrijken. In het ideale geval worden deze gegevens gedeeld tussen bedrijven en onderzoeksinstituten.

ACRONYMS

1

INTRODUCTION

CONSIDER the traffic situation of a pedestrian about to cross the road and a driver approaching the path of crossing in his or her vehicle (see Figure 1.1). They are two interacting traffic participants, perceiving each other, assessing the traffic situation, predicting their paths, and planning and negotiating their future actions to maneuver towards their goals safely. They visually perceive their environment to form a mental representation of the outside world and use experience to understand the surroundings.

The pedestrian is aware of his or her location, orientation, and intended path. By observing the approaching vehicle, he or she can estimate its location, orientation and velocity, predict its future path, assess whether the situation will become critical, and react accordingly (stop or continue crossing). As additional context information, the pedestrian observes the driver to assess the driver's awareness of the pedestrian, also

Figure 1.1: A traffic scenario of a pedestrian crossing the road as observed from an approaching vehicle. The left depicts the image from an on-board frontal-looking camera mounted behind the windshield of an intelligent vehicle. The center shows a cut-out of the pedestrian to ease judgment. Attentive human drivers extract a variety of context information besides the location of the pedestrian to interpret the situation. Just from a single image, a human driver can judge whether the pedestrian is aware of the approaching vehicle (head orientation), or is moving towards the other side of the road (legs apart). Similarly, when observing the image of a driver camera (right), one can judge that the driver is awake (eyes open), but not focusing on pedestrian (head and eyes pointing towards the building). If the driver has not focused on the pedestrian in the past, and with the information of the velocity of the vehicle, one can infer how critical the situation is and how likely a collision is without intervention, e.g., of an Autonomous Emergency Braking (AEB) system.

affecting the decision of whether it is safe to cross the road.

Likewise, the human driver observes the traffic scene, knows about the vehicle's location and velocity, and effortlessly localizes the pedestrian in relation to the road topology. The driver further infers important context information from the pedestrian's body and head pose, such as the crossing intention and the pedestrian's awareness of the driver/vehicle. This helps the driver to anticipate the pedestrian's behavior and react accordingly, e.g., by braking.

This everyday traffic situation seems effortless to humans and much of the reasoning happens subconsciously, based on innate cues such as gaze. How could an Advanced Driver-Assistance System (ADAS) perceive this situation, build a representation of the traffic scenario, extract useful information from the traffic participants, predict their future paths, and reason about situation criticality?

This thesis answers these questions by proposing a framework for an intelligent vehicle that intervenes in dangerous traffic situations by initiating emergency brakes or by emitting warnings to raise the awareness of the traffic participants. The framework presented in this thesis utilizes information from on-board sensors pointing-in (observing the driver) and pointing-out (observing the traffic scene, notably the pedestrian) and employs components from computer vision, pattern recognition, probabilistic reasoning, and state estimation. An overview of the framework will be given in Section 1.1.3 after further motivating the importance of road safety in Section 1.1.1.

1.1. MOTIVATION, SCOPE, AND CHALLENGES

Section 1.1.1 gives an overview of the current road safety situation and the behavioral key risk factors that cause accidents. They motivate special attention to pedestrians and drivers while they are interacting in traffic. Current ADAS already improve safety for pedestrians, as well as comfort and safety for drivers. While they consider pedestrians and driver individually, Section 1.1.2 suggests considering their mutual awareness to improve path prediction of both ego-vehicle and pedestrian to eventually increase road safety (see thesis scope in Section 1.1.3). The individual components like driver head pose estimation, as well as 3D person detection involve certain challenges, complemented by the challenges of their temporal integration for path prediction. They are described in Section 1.1.4.

1.1.1. ROAD SAFETY

More than 1.35 million people are killed yearly in traffic worldwide, and up to 50 million people are injured according to the World Health Organization (WHO) [135]. Pedestrians make up 23% of this number. About two thirds of serious crashes between vehicles and pedestrians occur while the pedestrian is crossing the road [36] and more than 32% occur on dedicated crossing locations with marked right-of-way (e.g., zebras, traffic lights) [37]. The main causes are distracted drivers, misinterpretation of the pedestrian's future action, obstructed view of the pedestrian and difficult lighting conditions [36]. This motivates to look into mutual awareness of driver and a potentially crossing pedestrian.

The WHO provides an interactive web visualization that helps to bring the dry figures into perspective and connects them to our emotions, see Figure 1.2: It depicts a map

of the world. In the center, it says: “A road user will die in 23 s” and counts down to 0 s. Afterward, it starts over. Yet another person died in traffic, and it raises the urge to do something about it. Road traffic injury is the leading cause of death for people aged between five and 29 years – while walking, cycling, or playing [135].

Figure 1.2: Screenshot of the interactive web visualization on traffic fatalities provided by the WHO, it provides a count down in seconds until the next death caused by traffic (statistically). Retrieved from https://extranet.who.int/roadsafety/death-on-the-roads on 2023-03-23, based on the data of [135].

There are inequalities across regions: Africa had 26.6 deaths per 100.000 population in 2016, compared to Europe with 9.3 deaths per 100.000 inhabitants. This inequality has further widened, as the number increased in Africa from 2013 to 2016, while it decreased further in Europe during that period [135].

There are behavioral key risk factors, such as (a) speeding, (b) drink-driving and drug-driving, (c) nonuse use of motorcycle helmets, seat-belts and child restraints, and (d) distracted driving [135]. These are legislatively being worked towards. In addition, building safer roads (e.g., better separation of traffic, especially vehicles vs. Vulnerable Road Users (VRUs)) and safer vehicles are important measures to improve road safety. Implementations of these measures vary between countries worldwide and human error is still a major cause of accidents [37].

These facts motivate this thesis to bring special attention to pedestrians and drivers while they are interacting in traffic. As artificial intelligence has found its way into series production vehicles, it opens new opportunities to understand driver-pedestrian interactions and help resolve critical situations.

1

1.1.2. DRIVER COMFORT, DRIVER ASSISTANCE AND AUTOMATED DRIVING

The last decades have already shown considerable improvements towards road safety and driver comfort, with many ADAS already available on the market. In addition, the aspiration of automated driving currently pushes investments and developments of vehicle manufacturers. The automated driving committee of the Society of Automotive Engineers (SAE) has published the J3016 standard, a taxonomy of driving automation levels with six levels of driving automation, incrementally shifting the distribution of responsibilities between driver and automation [114]. See Figure 1.3.

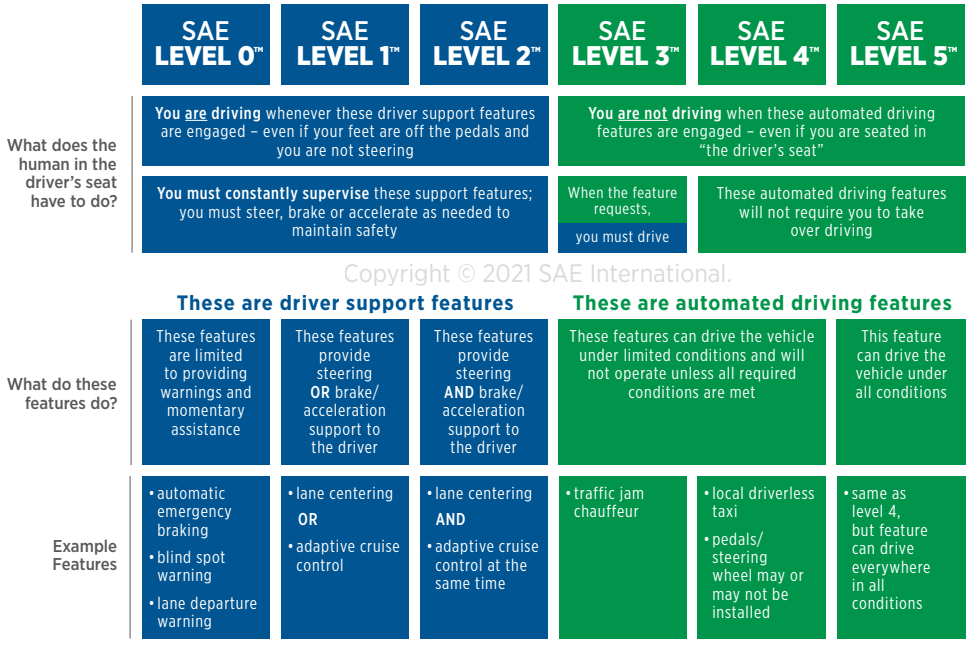


Figure 1.3: Visual chart of SAE J3016 levels of driving automation. It defines six levels of driving automation, from SAE level 0 (no automation), to SAE level 5 (full vehicle autonomy). Adopted from [114].

Up to including SAE level 4, the driver has to complement the automated system. Up to including SAE level 2, the driver must consistently supervise the support features, whereas in SAE level 3, the driver can engage in non-driving activities, but must take over the driving task on short notice when requested by the automation. SAE level 0 to SAE level 4 systems support only under specific conditions defined by the operational design domain (ODD), e.g. on-highway and during daytime. SAE level 0 features have shown a positive impact on driving safety statistics [25] by warning (e.g., forward collision warning, blind spot monitor, lane departure warning) or by intervention (e.g., automated emergency brake).

Regardless of SAE level of automation, early and robust localization of pedestrians is desirable, ideally with a large prediction horizon of the future path, respecting the cues from the pedestrian itself, but also from the infrastructure and other road users. When implementing pedestrian recognition systems, there is a trade-off to make between false

positives ("hallucinated" pedestrians) and false negatives ("overseen"/missed pedestrians). Ideally, both numbers should be close to zero and will decrease further over the next decades thanks to active ongoing research. Up to SAE level 2, the trade-off could be set towards fewer false positives while having more false negatives. This is possible because the driver is responsible for the driving task.

Towards pedestrian safety, Mercedes-Benz improved the active safety system PRE-SAFE®in 2013, which initiates emergency brakes if a dangerous traffic situation is found by camera-based pedestrian detection. In 2016, the Euro New Car Assessment Program (NCAP) has introduced tests for pedestrian protection systems, such as *AEB Pedestrians*, motivating further vehicle manufacturers to increase pedestrian safety.

When it comes to the driver, ADAS can make use of multiple cues to increase driving safety and comfort. Camera-based driver monitoring systems can detect fatigue/drowsiness [123], distraction/inattention [33], gestures [78], signs of being drunk [20], and readiness to take over from automated driving [30]. Camera-based driver head pose has been employed in on-market vehicles as early as 2007 (Toyota/Lexus) to estimate driver alertness. Cadillac (Super Cruise, 2018), BMW (Extended Traffic Jam Assistant, 2018), and Nissan (ProPilot, 2019) implement extended SAE level 2 capabilities and leverage a driver camera to assess the readiness of the driver to take over the task of driving. Mercedes-Benz's latest S-Class features a driver camera that monitors the driver's readiness to take over from automated driving mode on highways in an SAE level 3 system. This legally allows the driver to perform non-driving related tasks for up to 10 s under specific conditions. In addition, the latest S-Class features a volumetric heads-up display (HUD), an auto-stereoscopic 3D display, and multi-modal human-car interaction (e.g., the voice assistant inferring which window to open), each facilitated by head pose estimation.

Ideally, an intelligent vehicle with the driver in-the-loop knows where the driver puts his or her attention and can anticipate what the driver will do next. In ADAS, this can be used to lower false emergency brakes, e.g., in the scenario of a crossing pedestrian depicted in the beginning of this chapter, in the case where the driver is already aware of the pedestrian. With a deeper understanding of the driver, the scene around the vehicle, and how the road users interact comes a higher comfort and safety. To that end, this thesis addresses the interaction between the driver and a pedestrian and uses their mutual awareness to improve the prediction of their future paths.

1.1.3. THESIS SCOPE

The scope of this thesis is the path prediction and collision-risk estimation of ego-vehicle and a pedestrian based on features extracted from vehicle, driver and pedestrian. The focus is set on a pedestrian potentially crossing the road in front of the approaching ego-vehicle. The targeted application domain is an ADAS for SAE level 0–2 in traffic scenes with potentially crossing pedestrians. Integration into an intelligent vehicle is considered by using onboard sensors (as opposed to simulator data) as well as employing non-invasive sensors (as opposed to the driver or pedestrian wearing special equipment). Based on the predicted paths of ego-vehicle and pedestrian, collision risk is assessed and can be used in an adaptive AEB system. In this context, adaptive means that the trigger for AEB can be more relaxed if both road users are aware of each other.

Within the defined scope and application domain, this thesis presents a framework

that consists of multiple components: (a) camera-based driver head pose estimation with deep learning methods (pointing-in, Chapter 3 and Chapter 4), (b) camera- and lidar-based 3D person detection and orientation estimation with deep learning (pointing-out, Chapter 5), and (c) probabilistic path prediction of ego-vehicle and a potentially crossing pedestrian based on mutual awareness which is influenced by the driver head pose, vehicle position and speed, and features from the pedestrian (location, body orientation, head orientation) (Chapter 6).

A typical functional chain of an intelligent vehicle spans along the modules *sensors* > *perception* (incl. *sensor fusion*) > *behavior prediction* > *planning* > *control* [141]. This thesis contributes to multiple modules of the chain (see also Figure 1.4):

- *Sensors* provide measurements of the environment in a raw, machine-readable format. E.g., cameras measure photons and provide two-dimensional arrays of intensity values, and lidar sensors (actively) measure locations of multiple points with respect to the lidar sensor and provide them as point clouds. Both camera and lidar sensor data are leveraged in this thesis, whereas radar is out of scope.
- *Perception* extracts meaningful information from the raw sensor inputs, often with the use of pattern recognition methods. Specific to this thesis, driver head pose (3 degrees of freedom (DOF) translation and 3 DOF rotation) is extracted from single camera images, and 3D person location and body yaw rotation is extracted from a measurement pair of camera-image (forward-looking) and lidar point cloud. Multiple (noisy) per-sensor measurements are observed over multiple time steps to build an environment model. In this thesis, measurements of driver head pose, pedestrian location and orientation alongside ego-motion information of the ego-vehicle are fused in a probabilistic framework to estimate the motion dynamics of the road users.
- In *behavior prediction* future paths of ego-vehicle and other road users are estimated. The work of this thesis also touches upon the behavior prediction module by making probabilistic predictions of the future path of both ego-vehicle and a pedestrian.

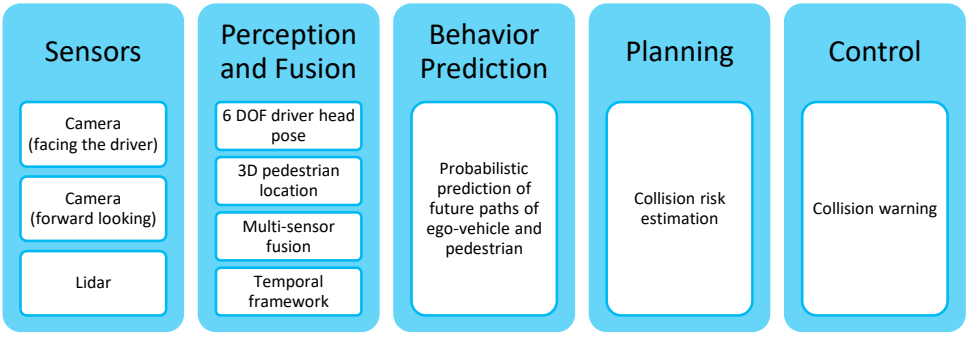


Figure 1.4: Typical functional chain of an intelligent vehicle (columns from left to right) and scope addressed by this thesis.

- *Planning* and *control* are components of an autonomous vehicle responsible for defining the future path of the ego-vehicle and using actuators to implement the path using motion control, e.g. apply brakes upon the decision of AEB. Collision risk estimation and warning of this thesis contribute to these modules.

Note that despite multiple pedestrians can be localized within the perception module, the path prediction experiments of this thesis focus on a single, potentially crossing pedestrian. This is to limit the complexity of interaction and keep data collection for optimization and evaluation feasible. Extensions to multiple pedestrians will be discussed in Section 6.6.

1.1.4. CHALLENGES

The intelligent vehicles domain brings certain challenges to overcome. There are challenges that are specific to each individual module mentioned in the last section, and there are challenges arising from composing the modules to a robust on-board system.

CHALLENGES TO VISION-BASED ALGORITHMS IN INTELLIGENT VEHICLES

Cameras mounted on-board a moving vehicle perceive the complex environment around and inside the vehicle (see Figure 1.5). Image sensors consist of a limited amount of pixels where the field of view is being projected to. For an object of fixed size, the size in pixels is inversely proportional to its distance, leading to small projections of far-away objects (see Figure 1.5a). Objects at the image border are only partly visible in the image (truncation, see Figure 1.5b).

Lighting conditions affect the exposure of the camera, e.g., low-light conditions (nighttime, dawn) increase the exposure time and can lead to motion blur for fast movements (see Figure 1.5c). Similarly, changes in lighting dynamics, such as entering/exiting tunnels, lead to delayed adaptations of the exposure time, causing temporal under/overexposure. Because the ego-vehicle is moving, the background continuously changes (as opposed to the surveillance domain with cameras mounted to static infrastructure) and is cluttered, making separation from objects of interest more challenging, also due to occlusions from other objects (see Figure 1.5b). Camera images might be disturbed by low sun, dirt, dust, rain, snow, or fog (see Figure 1.5d and Figure 1.5e).

1

(a) Low resolution.

(b) Truncation of person on left image border and hood of the ego-vehicle; occlusion by other pedestrians.

(c) Motion blur on left image border caused by long camera exposure time, high vehicle speed, and moving pedestrians.

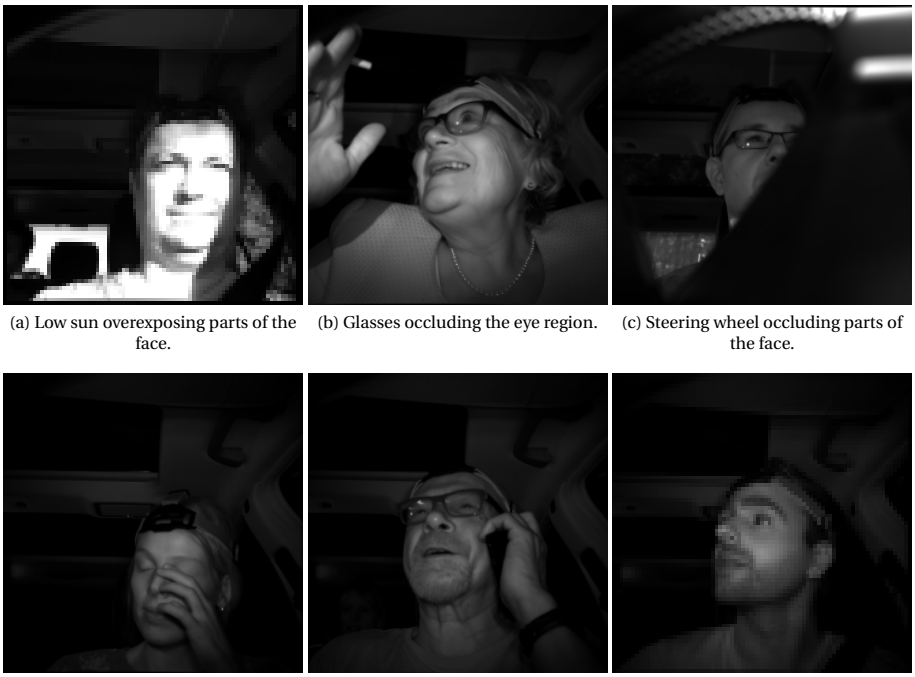
(d) Low sun causing artifacts on the image.

(e) Rain on the windshield locally blurring the image.

Figure 1.5: Examples of challenges of vision-based algorithms in intelligent vehicles perceiving the outside environment. Image source: EuroCity Persons dataset [12] with the author's permission.

CHALLENGES OF DRIVER HEAD POSE ESTIMATION

In-vehicle driver head pose estimation provides particular challenges in addition to those of general vision-based head pose estimation systems. The challenges include difficult illumination conditions (such as harsh sunlight covering parts of the face, see Figure 1.6a), occlusions (by worn objects such as glasses, see Figure 1.6b, or hands, see Figure 1.6d), but also due to the driving action (see Figure 1.6c), mobile phone use (see Figure 1.6e) and extreme head poses (see Figure 1.6f) imposed by naturalistic, complex driving scenarios while demanding a precise pose estimate and high availability in a non-invasive setting (no blinding illumination, no worn sensors). Different drivers vary widely in appearance, e.g., due to age, gender, ethnicity, or accessories. On the other hand, operating in-vehicle also provides advantages, such as a fixed perspective defined by the known extrinsic and intrinsic camera parameters, and the sparse number of faces simultaneously present within the cabin [76].



(a) Low sun overexposing parts of the face.

(b) Glasses occluding the eye region.

(c) Steering wheel occluding parts of the face.

(d) Hands occluding parts of the face.

(e) Hand-held mobile phone covering parts of the face.

(f) Out-of-plane rotations of the head.

Figure 1.6: Examples for challenges of vision-based driver observation algorithms.

When training and evaluating driver head pose algorithms, head pose distribution has to be considered: *Naturalistic* driving depicts mostly frontal head pose, especially during highway driving. Far-from-frontal head poses, e.g., during parking scenarios, occur less frequently.

CHALLENGES OF 3D PERSON DETECTION FROM CAMERA AND LIDAR

Detecting persons in urban traffic scenes poses further challenges in addition to the general challenges of vision-based algorithms mentioned above. Persons come in a wide variety of sizes, appearances, e.g., caused by clothing (different seasons, weather, time, personal style) or poses, see Figure 1.7. Persons are non-rigid: they can stand, walk, crouch, lie, jump, and depict complex articulations. Children, due to their body height are particularly difficult to detect in far distances, both for cameras (small projections), and for lidar sensors (few 3D points covering the child).

Figure 1.7: Examples for challenges induced by the intra-class variance of persons, causing different appearance (left) and poses, like sitting (right). Examples from the EuroCity Persons dataset [12] with the author's permission.

Detecting the 3D location of persons from camera images only has the challenge of missing depth information. Working only with data from a lidar sensor is challenging due to the sparse spatial distribution of point clouds at the distance. The lack of sensing texture makes it difficult to distinguish subtle features, such as head poses and overall body shape. Additionally, persons can be covered by fewer points due to external occlusion or self-occlusion compared to unobstructed cases.

Combining both modalities (camera images and lidar point clouds) introduces additional challenges, namely the need for precise calibration and synchronization. Intrinsic and extrinsic calibration parameters may change over time. In contrast to a camera, a lidar rotates while capturing the scene, so aligning point clouds of a moving vehicle demands for compensation of ego-motion.

CHALLENGES OF EGO-VEHICLE AND PEDESTRIAN PATH PREDICTION

Pedestrians are highly maneuverable; they can stop walking or change direction in an instant. This makes accurate path prediction a main challenge for intelligent vehicles. Road users dynamically interact with each other, which influences their future behavior, alongside environment factors. Low resolution of far-away pedestrians makes pedestrian head pose estimation imprecise. In turn, path prediction methods need to tolerate this lack of information/noise in the signal. Path prediction relies on observations of ego-vehicle movement, pedestrian location and pose, and driver head pose over time in a rigid, world-static coordinate frame. Therefore, the ego-centric perception has to be ego-motion compensated posing further challenges.

CHALLENGES OF SYSTEM INTEGRATION

Integration of the components into an intelligent vehicle poses additional demands. Processing should be performed in real-time with low latency, i.e., quickly reacting to new measurements. Sensors need to be and stay calibrated intrinsically and extrinsically: the optical projection parameters of each camera and the spatial transformation between all sensors and the ego-vehicle needs to be known. Sensors might decalibrate over time in a moving vehicle due to temperature changes and vibrations. This decalibration must be detected and compensated over time. Measurements of the sensors need to be synchronized, i.e. within the same time domain. Computation resources within an intelligent vehicle are limited: computation power is proportional to power consumption, which directly affects fuel economy, respectively battery range for electric vehicles.

1.2. OUTLINE AND CONTRIBUTIONS

The goal of this thesis is to increase traffic safety by improving path prediction of ego-vehicle and a potentially crossing pedestrian by leveraging mutual awareness of driver and pedestrian. To achieve this, this thesis contributes improved driver head-pose estimation, pedestrian localization and a probabilistic framework for path prediction based on features extracted from vehicle, driver and pedestrian.

Figure 1.8: Graphical outline of the chapters of this thesis (see also thesis cover for more visual context). Chapter 3 and 4 address interior-sensing of the ego-vehicle, specifically a driver head pose dataset and a 6 DOF head pose estimation method (camera frustum and driver head pose highlighted in yellow). Chapter 5 localizes the pedestrian in 3D (orange cuboid) based on exterior camera and lidar. Chapter 6 predicts paths of ego-vehicle (blue curves) and pedestrian (green uncertainty ellipses) based on driver head pose (yellow arrow), pedestrian location (orange cuboid) and pedestrian head pose (orange arrow).

Figure 1.8 and the cover of this thesis depict a graphical outline of the methodical chapters of this thesis. First, Chapter 2 presents previous work for the methodical chapters outlined above, namely, head pose datasets, head pose estimation methods, 3D person detection, and road user path prediction. Then Chapter 3 introduces a new large, naturalistic driver head pose dataset, motivated by the demand of deep neural networks with respect to dataset size and ground truth accuracy. The dataset is made available to the scientific community to encourage further research in this domain. Based on the dataset, Chapter 4 presents a novel driver head pose estimation method. Chapter 5 leverages a combination of camera and lidar sensors to estimate the 3D location and yaw rotation of persons surrounding the vehicle. Based on the output of the previous chapters, Chapter 6 presents a method that predicts the paths of the ego-vehicle and a pedestrian based on mutual awareness extracted from the head poses of both driver and pedestrian. In the final Chapter 7 overall conclusions are presented and future work is discussed.

The following subsections give a more detailed overview of the contributions of the methodical Chapters 3 – 6.

1.2.1. A LARGE-SCALE DRIVER HEAD POSE BENCHMARK

Developing head pose estimation algorithms based on machine learning demands for a large training set to optimize the large amount of parameters. Head pose, represented by 3 DOF of translation and 3 DOF of rotation, is difficult to manually annotate with sufficient precision. Chapter 3 introduces *DD-Pose*, the Daimler TU Delft Driver Head Pose Benchmark, a large-scale and diverse benchmark for image-based head pose estimation and driver analysis. It contains 330k measurements from multiple cameras acquired by an in-car setup during naturalistic drives. Large out-of-plane head rotations and occlusions are induced by complex driving scenarios, such as parking and driver-pedestrian interactions. Precise head pose annotations are obtained by a motion capture sensor and a novel calibration device. A high-resolution stereo driver camera is supplemented by a camera capturing the driver cabin. Chapter 3 is based on the work published in [109] (©2019 IEEE).

This chapter's main contributions are:

- The driver analysis benchmark from naturalistic driving scenarios features a broad distribution of head orientations and positions with an order of magnitude more samples of rare poses than comparable datasets. The dataset is made available for public benchmarking[1].
- The high-resolution stereo images allow for analysis of resolution, depth, and taking image context around faces into account.
- The supplemental camera of the driver cabin, combined with steering wheel and vehicle motion information, pave the way for holistic driver analysis, rather than head pose only.

[1] https://dd-pose-dataset.tudelft.nl

1.2.2. MONOCULAR DRIVER 6 DOF HEAD POSE ESTIMATION LEVERAGING CAMERA INTRINSICS

Chapter 4 presents *intrApose*, a novel method for continuous 6 DOF head pose estimation from a single camera image without prior detection or landmark localization. Using camera intrinsics alongside the intensity information is essential for accurate pose estimation. The proposed head pose estimation framework is crop-aware and scale-aware, i.e., it keeps poses estimated within image cut-outs consistent with the whole image. It employs a continuous, differentiable rotation representation that simplifies the overall architecture compared to existing methods. The method is validated on the dataset introduced in Chapter 3 and uses ablation studies to compare rotation and translation errors of intrinsics-aware and -agnostic methods, continuous and discontinuous rotation representations, and data sampling strategies. Chapter 4 is based on the work published in [108] (©2023 IEEE).

This chapter's main contributions are:

- It is observed that neglecting camera intrinsics (e.g., by using heuristics) introduces both rotation and translation errors that exceed reported rotation estimation errors. intrApose uses camera intrinsics consistently within the deep neural network and is crop-aware and scale-aware: poses estimated from bounding boxes within the overall image are converted to a consistent pose within the camera frame.
- This chapter borrows for the use in head pose estimation the continuous rotation representation SVDO$^+$ [69] which was used successfully in other domains.
- Using the challenging in-car driver head pose dataset introduced in Chapter 3, this chapter demonstrates that intrApose estimates translation and rotation more robustly compared to State-of-Art (SoA) methods, especially for extreme out-of-plane rotations.

1.2.3. DEEP END-TO-END 3D PERSON DETECTION FROM CAMERA AND LIDAR

Chapter 5 presents a method for 3D person detection from camera images and lidar point clouds in automotive scenes. The method comprises a deep neural network that estimates the 3D location, spatial extent, and yaw orientation of persons present in the scene. 3D anchor proposals are refined in two stages: a region proposal network and a subsequent detection network. For both input modalities high-level feature representations are learned from raw sensor data instead of being manually designed. To that end, the method presented in Chapter 5 uses Voxel Feature Encoders [146] to obtain point cloud features instead of widely used projection-based point cloud representations, thus allowing the network to learn to predict the location and extent of persons in an end-to-end manner. Experiments are conducted on the KITTI 3D object detection benchmark [43]. Chapter 5 is based on the work published in [110] (©2019 IEEE).

The main contributions of this chapter are threefold:

- A novel end-to-end deep learning-based method for 3D person detection using camera images and lidar point clouds is introduced. It does not rely on hand-crafted features.

- Various fusion schemes (early, late, and deep) and feature combinations (mean and concatenation) are compared.
- The method outperforms the prior SoA on the validation dataset of the KITTI 3D object detection benchmark [43].

1.2.4. DRIVER AND PEDESTRIAN MUTUAL AWARENESS FOR PATH PREDICTION AND COLLISION RISK ESTIMATION

Chapter 6 leverages the output of the methods presented in the former chapters, i.e., driver head pose and 3D person locations. It presents a novel method for vehicle-pedestrian path prediction that takes into account the awareness of the driver and the pedestrian towards each other. The method jointly models the paths of ego-vehicle and a pedestrian within a single Dynamic Bayesian Network (DBN). In this DBN, sub-graphs model the environment and entity-specific context cues of the vehicle and pedestrian (incl. awareness), which affect their future motion and allow to increase the prediction horizon. These sub-graphs share a latent state which models whether the ego-vehicle and pedestrian are on collision course; this accounts for a certain degree of motion coupling. The method is validated with real-world data obtained by on-board vehicle sensing (stereo vision, GNSS, and proprioceptive). Data consist of 93 vehicle and pedestrian encounters, spanning various awareness conditions and dynamic characteristics of the participants. Chapter 6 uses ablation studies to quantify the benefits of various components of the proposed DBN model for path prediction and collision risk estimation. Chapter 6 is based on the work published in [111] (©2022 IEEE).

The contributions are threefold:

- A method for joint path prediction and collision risk estimation of vehicle and pedestrian is presented that uses observed kinematics, mutual awareness, and environment cues. Joint awareness of driver and pedestrian towards each other has not been considered in previous work.
- An ablation study is provided analyzing the effect of various context cues on situations where an intervention of either road user is needed to avoid a collision.
- The proposed method is optimized and evaluated on newly collected real sensor data from a moving vehicle.

2

Previous Work

This chapter presents the previous work on the building blocks for driver-pedestrian mutual awareness and joint path prediction of ego-vehicle and pedestrian in the intelligent vehicles domain: camera-based driver head pose estimation including relevant datasets, 3D person detection and orientation estimation based on camera and lidar, and road user path prediction.

2.1. Driver Head Pose Estimation

This Section reviews different representations of rotations with a focus on applications within deep neural networks. Further, methods for image-based head pose estimation are surveyed.

2.1.1. Rotation Representations

Rotations in 3D space can be described by 3 degrees of freedom (DOF). There is an abundance of 3 DOF rotation representations, most prominently Euler angles, Tait-Bryan angles, rotation matrices, and quaternions. See Shuster *et al.* [122] for a survey and Table 2.1 for a tabular overview.

Euler angles and Tait-Bryan angles describe a rotation by three rotation components and an implicit or explicit convention of the order of axes the individual rotation components are applied on. For Euler angles the first and third rotation component are around the same axis while for Tait-Bryan angles, each rotation component refers to a dedicated coordinate axis. In addition, the rotation can be *extrinsic*, defining the rotation about axes of the original coordinate frame which is assumed to be motionless, or *intrinsic*, having the axes rotate along the chain of the three elemental rotations. This results in 12 different conventions for obtaining a well-defined rotation based on three given angles. Euler and Tait-Bryan angle components are restricted on a bound interval, thus introducing discontinuities (360° and 0° represent the same amount).

Rotation quaternions ($q \in \mathbb{H}$) are a compact 4-element representation allowing for efficient computation using quaternion algebra. Rotation quaternions suffer from the

Table 2.1: 3 DOF rotation representations within deep neural networks and the number of values (#val) they estimate. ⊥: representation is within SO(3) without post-processing and after each step of backpropagation. HPE: Applied to head pose estimation. ↻: representation is continuous in accordance with the definition of Zhou *et al.* [144].

Representation	#val	Method	⊥	HPE	↻
YPR [145]	3	Euler / Tait–Bryan angles	✓	✓	-
Rotation vector [4]	3	(rotvec). Compact axis-angle	✓	✓	-
Axis-Angle [144]	4	3-vector for axis, scalar angle	-	-	-
Quaternion [54]	4	Quaternion + normalization	-	✓	-
Ortho5D [144]	5	Stereographic projection	✓	-	✓
Ortho6D [144], [52]	6	Gram-Schmidt process	✓	✓	✓
M [144]	9	3x3 matrix (unconstrained)	-	-	-
SVD-inf [16, 69]	9	SVD (inference) on M, ortho loss [16]	-	✓	-
SVDO$^+$ [69]	9	*Differentiable* SVD (training) on M	✓	-	✓

antipodal problem making q and $-q$ represent the same rotation [122].

A rotation matrix R ($R \in SO(3) \subset \mathbb{R}^{3\times3}$, $RR^T = I$, $\det(R) = +1$) maps an orthonormal basis in $\mathbb{R}^3$ to another orthonormal basis in $\mathbb{R}^3$, spanned by the three columns of R. SO(3) is the special orthogonal group containing all rotations in 3D.

There are less frequently used representations, such as axis-angle (axis $a \in \mathbb{R}^3, ||a||_2 = 1$, angle $\theta \in [0, 2\pi]$), and rotation vectors (rotvec; $r \in \mathbb{R}^3$, with angle $\theta = ||r||_2$).

The above representations have the drawback of being discontinuous, which are less suitable for learning, by leading to higher errors or slower convergence [144]. Recently, rotation representations have been proposed to overcome these drawbacks. Zhou *et al.* [144] proved that in the three-dimensional space any rotation representation with less than five dimensions is discontinuous and thus harder to approximate by a neural network. Zhou *et al.* [144] construct an Ortho5D and an Ortho6D representation which are both continuous. Out of the 5 (6) values, a rotation matrix $\in SO(3)$ is built using a stereographic projection (a Gram-Schmidt process). Levinson *et al.* [69] explore the viability of integrating symmetric orthogonalization SVDO$^+$ (based on singlar value decomposition (SVD)) directly into the neural network following an unconstrained intermediate representation of nine values (a degenerate rotation matrix M). SVDO$^+$ is continuous and differentiable, thus suitable within deep neural networks.

2.1.2. HEAD POSE ESTIMATION

Head pose estimation from images has been a popular topic for decades and can be categorized from different perspectives, i.e., the *methodical* perspective, the *I/O (input/output)* perspective, and the *application* perspective.

METHODICAL PERSPECTIVE

Both surveys of Murphy-Chutorian *et al.* [81] and Abate *et al.* [1] use a high-level method-based categorization: template-based methods, subspace-based methods, feature-based

methods, and regression-based methods.

Template-based methods estimate head pose by matching appearance templates, i.e. by comparing test images to a set of exemplars with known pose. *Subspace-based methods* map the input space (e.g. image intensities) to a head pose manifold. E.g., Derkach *et al.* [31] use tensor decomposition to model a non-linear manifold of 3D head poses. *Feature-based methods* make use of an intermediate geometric representation of the face. E.g., Baltrusaitis *et al.* [7] localize facial landmarks by a constrained local model (CLM) and estimate head pose by a successive generalized adaptive view-based appearance model. Tran *et al.* [129] fit a 3D morphable model to the head which implicitly encodes head pose. Chang *et al.* [18], point out several drawbacks of facial landmark locations: they (a) are ill-defined, therefore vary in interpretation of the annotator, (b) represent facial contours, therefore change with a different viewpoint, and (c) become occluded depending on viewpoint. This introduces certain errors in head pose estimates based on facial landmark locations.

Regression-based methods learn a (non-linear) functional mapping from input data to the head pose parameter space. There is an abundance of work within this domain, so some examples representative of the concept are pointed out and the reader is referred to the comprehensive survey on deep regression by Lathuiliere *et al.* [66]. The methods typically perform regression using a neural network, though other regression models have been applied. Neural network-based regression models consist of a CNN-based backbone for feature extraction and a prediction head. Regressing a discontinuous rotation representation has an impact on the architecture. One prominent scheme is coarse-and-fine/ordinal regression, where coarse bins are classified in addition to continuous regression values [16, 55, 113, 133, 139]. Zhou *et al.* [145] address the discontinuities of large Euler angles by a wrapped loss. Schwarz *et al.*[119] choose quaternions and use a regularization term to keep quaternion elements small. Hsu *et al.* [54] additionally propose losses explicitly dealing with the inter-dependence of certain quaternion elements and the independence of others. Albiero *et al.* [4] use a rotation vector representation. Their method img2pose estimates the delta from a normalized pose (zero-mean, unit standard deviation) to increase robustness. Further, img2pose employs a calibration point loss which uses a set of head-static 3D points (e.g., 3D head landmarks) and compares the projected points of ground-truth and predicted pose.

Lately, methods have tackled the constraint of orthogonality. Cao *et al.* [16] propose to estimate the basis vectors or the rotation matrix and use a loss to keep the basis vector close to orthogonal. Yet, SVD is needed to create an orthonormal rotation representation. Zhou *et al.* [144] have proposed the continuous Ortho6D representation (see Section 2.1.1). Hempel *et al.* [52] applied it to a head rotation estimation in a simple deep neural network estimating six values, and employ the Ortho6D representation. The method does not estimate head translation.

I/O (INPUT/OUTPUT) PERSPECTIVE

A taxonomy orthogonal to the above structure follows the available input modality (e.g., intensity, depth [9, 39], optical-flow or a combination of those), and whether a single measurement in time is being used or multiple consecutive frames (e.g., tracking [8, 121], RNNs [47]). Another disambiguation is about which of the 3 DOF rotation parameters

are being estimated, e.g. from a single yaw angle [32, 99] up to 3 DOF head rotation parameters. Finally, methods can be distinguished by whether they estimate (up to 3 DOF) translation alongside rotation, and whether a preprocessing step is needed before the pose estimation, such as a face bounding box detection. This thesis focuses on single-image, intensity-based methods estimating continuous, full rotation (3 DOF) and translation (3 DOF).

APPLICATION PERSPECTIVE

Based on the application domain, different challenges/requirements arise that need to be addressed by the method. E.g., surveillance applications typically have to deal with low-resolution and tolerate larger rotation errors. A widely used application is head pose estimation within generic images. Generic images are typically easy to obtain (e.g., collected from the internet), but lack other information, such as camera intrinsics. Within this category one recent method is img2pose [4], a Faster R-CNN-based [105] head pose estimation method which estimates full 6 DOF head pose without prior face or landmark detection. The method regresses bounding boxes out of which features are being pooled for a prediction head. The prediction head regresses a discontinuous rotation vector and a translation vector for each bounding box. The prediction head loses context by being presented with a cut-out of the whole image. Therefore, the bounding-box-local pose is converted to an image-global pose using scaling heuristics. img2pose implicitly assumes a fixed focal length for all input images, leading to erroneous head pose estimates if the assumption fails (i.e., with images depicting a different field-of-view). The ground truth pose used for training and evaluation is obtained using the same focal length assumption and is thus biased.

Another domain is in-vehicle applications. Most approaches focus on head rotation from depth data from structured infrared light, such as Borghi *et al.* [9], Schwarz *et al.* [120], or Venturelli *et al.* [131], while Ahn *et al.* [2], Firintepe *et al.* [41] and Schwarz *et al.* [119] leverage intensity images to estimate head rotation. Out of the aforementioned methods, only Schwarz *et al.* [120] estimate head translation in addition to rotation, yet only from depth data.

See Table 2.2 for the nearest neighbors of the method presented in Chapter 4.

Table 2.2: Related methods for intensity-based head pose estimation with their translation and rotation representations. Crop-aware: whether the pose is consistent with the image crop. In-vehicle: whether the method has been applied to driver head pose estimation. Intr.-aware: whether the method leverages camera intrinsics. Note that the top four methods do not estimate head translation, yet are of interest due to their rotation representation.

Name	Method	Translation	Rotation Representation	Crop-aware	In-vehicle	Intr.-aware
TriNet [16]	Coarse-to-fine, SVD-inf	-	R	-	-	-
QuatNet [54]	Quaternion, coarse-and-fine	-	Quaternion	-	-	-
WHENet [145]	Coarse-and-fine, address Euler discontinuities	-	Euler (YPR)	-	-	-
6DRepNet [52]	Ortho6D representation of [144]	-	R	-	-	-
img2pose [4]	Faster R-CNN, crop-invariant proposals	XYZ	Rotation vector	✓	-	-
intraApose (Chapter 4)	**Crop-invariant proposals, SVDO+**	**XYZ**	**R**	✓	✓	✓

2.2. HEAD POSE DATASETS

There is an abundance of publicly available camera-based head pose datasets dating back nearly two decades [5, 6, 39, 47, 64, 75, 76, 84, 89, 117, 119, 140] (see Table 2.3).

Head pose datasets can be categorized by different aspects, such as *imaging characteristics*, *data diversity*, *acquisition scenario*, *annotation type*, and *annotation technique*. These aspects play an important role on whether and how the dataset identifies challenges of the head pose estimation task.

Imaging characteristics relate to the image resolution, number of cameras, bit depth, frame rate, modality (RGB, grayscale, depth, infrared), geometric setup and field of view.

Data diversity incorporates aspects such as the number of subjects, the distribution of age, gender, ethnicity, facial expressions, occlusions (e.g. glasses, hands, facial hair) and head pose angles. Data diversity is essential to training and evaluating robust estimation models.

Acquisition scenario covers the circumstances under which the acquisition of the head pose takes place. The most important distinction is between in-laboratory [5, 6, 39, 47, 117, 140] vs. in-the-wild [75, 76, 84, 89, 119] acquisition. While the former restricts the data by defining a rather well-defined, static environment, the latter offers more variety through being acquired in unconstrained environments such as outside, thus covering many challenges like differing illumination and variable background. Head movement can be *staged* by following a predefined path or can be *naturalistic* by capturing head movement while the subject performs a different task, such as driving a car.

Annotation type describes what meta-information, such as head pose, comes alongside the image data and how it is represented. *Head pose* is defined by a full 6 DOF transformation from the camera coordinate system to the head coordinate system, covering 3 DOF for translation and 3 DOF in rotation. Head pose datasets differ in how many of those DOF are provided alongside the images, i.e. whether only a subset of the translation and rotation parameters is given. Ultimately, annotation types differ in their granularity of sampling the DOF space: there are discrete annotation types that classify a finite set of head poses, and there are continuous annotation types that offer head pose annotations on a continuous scale for all DOF.

There are different *annotation techniques* for obtaining the head pose annotation accompanying each image. The annotation technique has a large impact on data quality. It can be categorized into manual annotations vs. automatic annotations. For manual annotations, human experts annotate the image data according to a label specification [117]. Automatic annotations can be divided into data-based annotations, computed by algorithms on the image data [39, 140], and sensor-based annotations, which in turn use an additional hardware sensor for obtaining the head pose for each image [5, 6, 119].

Manual annotations do not need additional hardware but are prone to introduce errors and biases. E.g., a human annotator can only annotate in the image plane, thus needs to guess the distance part of the translation of the head [75, 76]. There is also inter-annotator variability through different interpretation of the same scene. Additionally, as manual annotations consume human time, their cost scales linearly with the amount of data to be annotated.

Automatic annotations based on algorithms computing the annotations from the image data are fast to obtain but induce systematic errors of the underlying algorithm

and will not allow disambiguating between annotation errors and errors induced by the method under test.

Automatic annotations based on sensors make use of additional reference sensors during the data acquisition process. The reference sensor measurements should be calibrated to the head coordinate system and calibrated and synchronized to the camera images. There are different types of reference sensors that differ in their measurement method. Among those are electromagnetic sensors [5, 6], inertial sensors, vision-based sensors, 3D scanners [117], optical marker tracking sensors [119], and hybrid combinations of them. An optimal reference sensor for head pose estimation should be accurate, free of drift, robust to disturbance, and measure all 6 DOF on a continuous scale.

From the aspects mentioned above, this thesis focuses on datasets with continuous head pose annotations for all 6 DOF which offer naturalistic scenarios and a large data diversity.

Section 2.1.2 showed that many recent models for classification and regression tasks are based on deep convolutional neural networks.Their high model complexity demands for a very large number of training examples. Therefore, this thesis also focuses on large datasets in terms of number of images.

An overview of currently available datasets is given in Table 2.3. Respective example data can be found in Figure 2.1.

Table 2.3: 2D/3D face datasets with continuous head pose annotations.

Dataset	GT	Year	#Cams x w x h	#Images	#Subjects f/m	Head pose	Reference	Scenarios
Bosphorus [117]	3D	2008	1x1600x1200	5 k	45/60	relative	guided	choreographed facial expressions
ICT-3DHP [6]	3D	2012	1x640x480	1 k	6/4	relative	magnetic	choreographed large rotations
Biwi Kinect [39]	3D	2013	1x640x480	16 k	6/14	relative	guided, ICP	choreographed large rotations (yaw, pitch)
gi4e hpdb [5]	2D	2016	1x1280x720	36 k	4/6	relative	magnetic	choreographed large rotations
SynHead [47]	3D	2017	1x400x400	511 k	5/5	absolute	synthetic data	70 different motion tracks
UbiPose [140]	3D	2018	1x1920x1080	14 k	22[c]	absolute	3DMM	service desk interactions
RS-DMV [84]	2D	2010	1x960x480	13 k	6[c]	N/A	N/A	naturalistic driving
Lisa-P [75]	2D	2012	1x640x480	200 k	14[c]	relative	POS [29]	naturalistic driving, choreographed large yaw
NDS HPV [89]	2D	2015	1x720x480	2 PB[d]	>3100[c]	N/A	N/A	naturalistic driving
VIVA [76]	2D	2016	1x*544	1 k	N/A	relative	POS [29]	naturalistic driving
DriveAHead [119]	3D	2018	1x512x424[a]	1 M	4/16	absolute	mo-cap	naturalistic driving, parking
DD-Pose (Chapter 3)[b]	3D	2019	2x2048x2048	2x330 k	6/21	absolute	mo-cap	naturalistic driving, large rotations and translations

[a] only head image crops provided. Mean size 25x50

[b] additional data streams recorded: front facing camera, interior camera facing driver from the rear right

[c] female/male ratio not provided by the authors

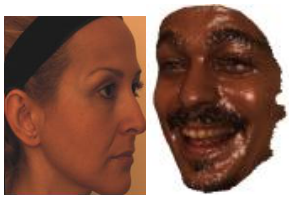
[d] number of images not provided. Assumed to be $>10^9$

The datasets are subdivided by their acquisition scenario into two groups, namely generic head pose datasets vs. driving head pose datasets. The latter come with desirable properties such as naturalistic scenarios, a large data diversity and challenging imaging characteristics.

2.2.1. GENERIC HEAD POSE DATASETS

Bosphorus [117] contains 5k high resolution face scans from 105 different subjects. The 3D scans are obtained by a commercial structured-light based 3D digitizer. It offers 13 discrete head pose annotations with different facial expressions and occlusions.

ICT-3DHP [6] provides 1400 images and depth data from 10 subjects acquired with a

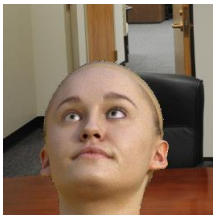


(a) Bosphorus [117]

(b) Biwi Kinect [39]

(c) ICT-3DHP [6]

(d) gi4e hpdb [5]



(e) SynHead [47]

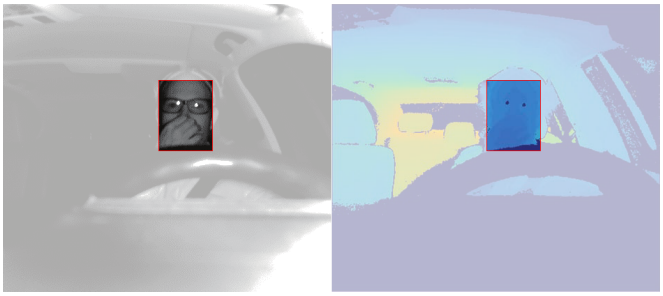
(f) UbiPose [140]

(g) RS-DMV [84]

(h) Lisa-P [75]

(i) NDS HPV [89]

(j) VIVA [76]



(k) DriveAHead [119]. Content within red box provided. Left: infrared camera image. Right: Depth image.

Figure 2.1: Example data of the investigated 2D/3D head pose datasets. The datasets differ in many aspects, such as sensor modalities (RGB, IR, depth), in-lab vs. synthetic vs. naturalistic driving, precision of head pose annotation and resolution.

Kinect v1 sensor. 6 DOF head pose annotations are measured by a magnetic reference sensor. The authors do not detail on whether calibration and synchronization of the reference sensor measurements to the camera images is performed.

Biwi Kinect [39] consists of 16k VGA images and depth data from 20 subjects depicting the upper body. The data were acquired by a Kinect v1 sensor. 6 DOF head pose annotations are provided by fitting user-specific 3D templates on depth data, which has limitations when occlusions are present. As it is recorded in a laboratory environment, it provides a uniform and static background.

gi4e hpdb [5] contains 36k images from 10 subjects recorded with a webcam in an in-laboratory environment. Head pose annotations are given in 6 DOF using a magnetic reference sensor. All transformations and camera intrinsics are provided. Head pose annotations are given relative to an initial subjective frontal pose of the subject.

SynHead [47] contains 511k synthetic images from 10 head models and 70 motion tracks. The rendered head models are composed with random background images, providing indoor/office scenery. As this is a generative method for data synthesis, head pose annotations are very accurate. Making use of 10 head models provides little diversity of human facial expressions.

UbiPose [140] features natural role played interactions with 10k frames obtained by a Kinect v2 sensor. 22 subjects are recorded. Head pose was annotated automatically based on the raw footage using initial facial landmark annotations and fitting a 3D morphable model. Annotations not fitting the data were pruned by human annotators. Subjects were captured from a relatively large distance.

2.2.2. HEAD POSE DATASETS IN THE AUTOMOTIVE CONTEXT

RS-DMV [84] contains 13k images from six subjects captured in naturalistic outdoor and simulator scenarios. Head pose annotations are not provided.

Lisa-P [75] offers 200k images from 14 subjects with a resolution of 640x480. Head rotation annotations are obtained by using the Pose from Orthography and Scaling (POS) algorithm [29] on manually labeled facial landmarks. By using an orthographic projection, this approach only allows for approximate position and rotation estimates.

NDS-HPV [89] contains 2 PB (peta byte) of highly compressed, low resolution images from a naturalistic driving study. It contains images of over 3100 subjects collected over a period of over two years. Head pose annotations are not provided, thus restricting its use to qualitative analysis.

The VIVA head pose estimation benchmark [76] is a test set consisting of images with 607 faces, out of which 323 are partially occluded. The naturalistic driving images were selected both from research vehicle recordings and YouTube videos to display harsh lighting conditions and facial occlusions. The head pose annotations of the test dataset are not released, but evaluation is possible by submitting hypotheses through a benchmark website. No training images are provided.

DriveAHead [119] features 1M images and depth information acquired by a Kinect v2 sensor during naturalistic driving. 20 different subjects appear in the recordings. Images were collected with a resolution of 512x424 pixels. 6 DOF continuous head pose annotations are obtained by a motion capture system which measures the pose of a marker fixated at the back of the subject's head. The coordinate transformation between

the head mounted marker coordinate system and the head coordinate system is calibrated per-subject by measuring the position of eight facial landmarks of the face of each subject after fixating the head-mounted marker. The transformation between the reference sensor coordinate system and the camera coordinate systems are known, although the calibration process is not described. Alongside, per-image annotations for occlusions and whether the subjects wears glasses or sunglasses is provided.

The large number of image samples enables training of deep convolutional neural networks for head pose estimation. Parking maneuvers and driving on a highway and through a small town results in naturalistic head movements, thus providing distributions of head rotation angles and head positions which are typical for naturalistic drives.

As no intrinsic camera parameters are provided, 3D points in the camera coordinate system cannot be projected into the image space. Consequently, both head position and rotation estimation methods have to implicitly adapt to the specific dataset. DriveAHead provides cut-outs of faces with a mean inter-pupil distance of 35 pixels, thus targeting on methods for low-resolution head pose estimation.

2.3. 3D Person Detection

This section first reviews deep learning based object detection methods, followed by methods for 3D object detection applied to detection of Vulnerable Road Users (VRUs) based on camera-only, lidar-only and both modalities combined.

2.3.1. Deep Learning based Object Detection

Deep neural networks, specifically convolutional neural networks (CNNs) have been shown to be very accurate in many image recognition tasks such as image classification [62], object detection and especially person detection [12]. A large number of artificial neurons stacked in several layers create a neural network. It transforms the input datum into an output datum which represents the task to solve. Within each layer, a more high-level representation of the input is encoded.

For the task of object detection, there are two different approaches, namely *two-stage* and *single-stage* object detectors.

Two-stage Object Detection

Two-stage object detection architectures consist of a region proposal stage and a proposal classification stage. The region proposal stage generates proposals which are to be classified by the proposal classification stage. Using a fast region proposal stage allows to keep inference time low, while still maintaining a high accuracy. Popular methods adopting this approach are Region-based Convolutional Neural Networks (R-CNN) [46], Fast R-CNN [45], Faster R-CNN [105] and Region-based Fully Convolutional Network (RFCN) [27].

Single-stage Object Detection

Single-stage object detection architectures perform the detection task in one forward pass through the network. You Only Look Once (YOLO) [100] and Single-shot multibox detector (SSD) [74] are prominent representatives of single-stage object detectors.

YOLO and its improved derivatives [101, 102] divide an input image into a grid. Each grid cell predicts a fixed number of bounding boxes with an associated confidence score. Each bounding box is classified. The resulting detections are obtained by a non-maximum suppression (NMS).

SSD uses a base CNN to extract feature maps at different layers. Each layer produces detection proposals based on default bounding boxes associated to each feature map. This allows for specialized classification of objects in various sizes.

2.3.2. 3D PERSON DETECTION IN THE AUTOMOTIVE CONTEXT

3D person detection in the automotive context can be performed on measurements acquired by different on-board sensors such as cameras, radar and lidar. Methods can perform either on a single modality or a combination of sensor inputs. This section focuses on camera-only methods, lidar-only methods and methods working on both modalities. Radar-based methods are out-of-scope for this thesis. Palffy *et al.* [87] presents an exemplary method leveraging radar. Qian *et al.* [97] present a recent survey on 3D object detection for autonomous driving.

CAMERA-BASED METHODS

[80] estimates the 3D pose of objects from a single image. A State-of-Art (SoA) 2D detector is used to obtain 2D bounding boxes of the objects. A CNN estimates the 3D pose of the object while considering projective geometry constraints. 3DOP [22] generates 3D object proposals from stereo-based depth information. An energy function is formulated to exploit different features such as prior object size, free-space, and point densities inside the bounding box. The 3D object proposals are then scored by a CNN. In contrast, Mono3D [21] creates 3D object proposals from monocular images by exploiting constraints such as objects residing on the ground plane. Proposals are scored by semantic information, context, as well as shape features and location priors. [17] introduces Deep MANTA, a CNN which estimates 2D bounding boxes and vehicle part locations, along with visibility and a 3D CAD template. The pose in terms of location and orientation is recovered by using a 2D/3D point mapping.

LIDAR-BASED METHODS

In contrast to images, point clouds are inherently unordered and have a varying size. To overcome the this issue, different representations have emerged: bird eye view (BEV) [23, 124], sensor-view [70], mixed 2.5D BEV images [65], and voxel grids [146]. These representations allow for transplanting image-based methods to point clouds, specifically CNNs [70, 124].

In [70], an image-like 2D point map representation is used on which a CNN is employed for 3D object detection. Complex-YOLO [124] expands the YOLOv2 CNN [101] and applies it on a BEV representation of the point cloud to detect 3D objects. These methods rely on *designing* a good representation of point clouds.

In contrast, there there are methods which *learn* features from point clouds without a strongly enforced intermediate representation [19, 65, 96, 138, 146]. PointNet [19] presents a permutation-invariant deep neural network which learns global features from unordered point clouds. The method is applied to 3D part segmentation and point-wise

semantic segmentation. PointPillars [65] uses the features from PointNet in order to learn a feature representation on vertical columns (pillars). This allows for an image-like representation on which an SSD-based detection network is applied for 3D object detection.

Voxelnet [146] avoids using hand-crafted features by partitioning the input space into equally-sized voxels. The group of points within each voxel is transformed into a unified feature representation through voxel feature encoding (VFE) layers. A VFE layer combines point-wise features with a locally aggregated feature. Stacking VFE layers allows for learning higher level features. The resulting high-dimensional volumetric representation is used in a region proposal network framework to estimate the 3D location of objects.

CAMERA- AND LIDAR-BASED METHODS

There are multi-modal fusion methods which combine camera images and lidar point clouds. At the cost of needing a well-synchronized and calibrated sensor setup, benefits from each modality can be leveraged. E.g. small and far-away objects can be visible in the camera image while they may not have a lidar measurement.

Frustum PointNets [95] uses a SoA 2D object detector on camera images to obtain points which reside in the object's frustum. Then, 3D object instance segmentation and bounding box regression is performed on point features extracted with the method of PointNet [146].

Multi-View 3D (MV3D) [23] and Aggregate View Object Detection (AVOD) [63] are both sensor fusion methods, i.e. they take input from both camera images and lidar point clouds, extract features from each modality, fuse the features and consequently perform 3D bounding box regression.

MV3D [23] represents point clouds in a front view and a BEV image. Along with the camera image, convolutional layers are applied for high-level feature representation. A region proposal network creates view-specific feature crops from the BEV. The per-modality region-based features are fused either *early*, *late*, or in a *deep* fusion scheme. 3D bounding box regression is performed to obtain the object's 3D position.

AVOD [63] provides a similar architecture as MV3D. However, it fuses features from the individual sensors earlier in the region proposal network in order to capture smaller objects. Furthermore, AVOD uses the camera image and BEV input only, while still achieving a higher accuracy on the KITTI 3D detection benchmark.

Both MV3D and AVOD present multi-modal architectures for 3D object detection from camera and lidar. However, the methods rely on a dense image-like feature representation of the inherently sparse point cloud.

2.4. ROAD USER PATH PREDICTION

Road user path prediction has attracted a lot of attention in recent years, see surveys regarding the ego-vehicle [68] and Vulnerable Road Users [106, 112]. Path prediction methods require positions as input. Ground plane positions relative to a vehicle coordinate system can be obtained from detections in various sensors (e.g. camera [12], radar [86], lidar [126], or a combination thereof [110, 126]). See also Section 2.3 for more previous work on 3D person detection. If ground plane positions relative to a global coordinate system are needed (e.g., this thesis), then vehicle ego-motion compensation is necessary

as an additional pre-processing step. For this, a combination of GNNS, INS, and vehicle proprioceptive sensing can be used. Following sub-sections focus on context cues and motion models used for path prediction.

2.4.1. CONTEXT CUES FOR PATH PREDICTION

In the most rudimentary form, cues for path prediction consist of point kinematics, i.e. positions and velocities of the relevant object. It has however been well established that the use of additional "context" cues can improve path prediction performance [112]. These can be categorized into object cues, and static and dynamic environment cues.

Object context cues refer to cues pertaining to the object of interest itself. For example, Keller and Gavrila [58] improve pedestrian path prediction by using dense optical flow features extracted from a pedestrian bounding box. Kooij *et al.* [60] use relative head orientation as a "proxy" for the pedestrian's awareness of the oncoming ego-vehicle while crossing. Kooij *et al.* [61] and Pool *et al.* [93] incorporate the arm gesture of a cyclist to predict its turn at an intersection. Quintero *et al.* [98] recover a full 3D articulated pose of a pedestrian to better predict crossing action.

Object context cues can also refer to properties derived from the driver of the ego-vehicle, when interested in predicting the future ego-vehicle path. Typical such cues are driver head orientation or gaze, or performed driver actions, as inferred from accelerator pedal position, braking force and steering wheel angle. For example, Roth *et al.* [107] employ driver head pose to capture the driver's awareness of a crossing pedestrian.

Static environment context cues refer to elements of the static traffic infrastructure which will likely influence road user motion, such as road topology [93, 94], road markings and traffic lights [130].

Dynamic environment context cues capture the presence and motion properties of other road users (including that of the ego-vehicle itself) that may influence the target road user's behavior, i.e. to avoid hazards or to minimize hindrance. For example, [60, 61, 83, 90, 107] use basic kinematics properties, such as relative distances and velocities, and the expected point of closest approach.

2.4.2. MOTION MODELS

Models for human motion path estimation can be subdivided into physics-based, pattern-based and planning-based methods [112]. As motivated earlier, this thesis focuses on physics-based methods, which represent motion by explicitly defined dynamic equations of one or more underlying dynamical models. Simple motion dynamics can be modeled by Linear Dynamical Systems (LDSes), which commonly assume a linear relationship between states and measurements with Gaussian noise. Under these assumptions, the Kalman Filter (KF) [134] is an optimal filtering algorithm, which has been widely applied for pedestrian and vehicle tracking [68, 118].

In the scope of collision analysis, motion models play a role for predicting paths of targets such as a potentially crossing pedestrian and the ego-vehicle. The probabilistic models described here allow to extrapolate observed behaviors into the future while accounting for uncertainties in the assumed dynamics and observations.

Since traffic behavior may change at any time, a common approach is to treat the complex dynamics by switching between or combining multiple motion models at each

prediction step, e.g., by using a Switching Linear Dynamical System (SLDS). SLDSes can be extended by dynamical models to incorporate contextual cues for path prediction [61, 98]. Li *et al.* [72] combine the path prediction output of Kooij *et al.* [61] with a sequence-to-sequence path generation method to leverage the complementary advantages of hand-crafted models and data-driven methods.

Different methods have been introduced to predict the paths of multiple interacting road users, e.g., Social Force models for human-human interactions [51]. For pedestrian-vehicle encounters, e.g., Kooij *et al.* [61] assume that the vehicle does not change motion dynamics, while Braeuchle *et al.* [10] use a Bayesian Network to find an appropriate vehicle motion model which minimizes pedestrian injury risk. The pedestrian motion model is fixed based on initial velocity. Gupta *et al.* [49] simulate actions (speed up, slow down) of a self-driving vehicle within a negotiation cycle with a crossing/yielding pedestrian to optimize traffic throughput.

2.5. COLLISION RISK PREDICTION

Collision risk prediction can be categorized into physical model-based and data-driven methods [26]. The latter estimate collision risk metrics based on training data. Physical model-based methods incorporate physical knowledge and can further be subdivided into single-behavior threat metrics (SBTM), optimization-based methods, formal methods, and probabilistic approaches. However, these categories can partially overlap [26]. Söntges *et al.* [127] present an SBTM method by computing time-to-react from over-approximating reachable sets. DeNicolao *et al.* [28] directly estimate collision risk based on ego-vehicle motion and a random-walk-based pedestrian motion simulation. Collision risk is precomputed by simulation of pedestrian crossing and looked up during inference based on ego-vehicle motion model parameters and relative position.

Given predictive distributions, collision risk can be obtained by analytic [10] or discrete [14] integration. Bräuchle *et al.* [10] use a compound car-pedestrian geometric model to infer a joint spatial probability distribution. Collision risk is estimated by integration over predicted distributions for all time steps. The method of Brouwer *et al.* [14] fuses predicted object occurrences from four pedestrian motion models in a probabilistic fusion grid. Collision risk is estimated by summation over all grid cells inside the collision corridor. Roth *et al.* [107] estimate a joint spatial distribution by moment matching of predicted distributions for vehicle and pedestrian. A collision risk is calculated by integrating the joint spatial distribution over the collision area, defined by all possible intersections between vehicle and pedestrian locations.

3

A LARGE-SCALE DRIVER HEAD POSE BENCHMARK

This chapter introduces a new driver head pose dataset that is important for the development of the head pose estimation method of Chapter 4.

3.1. OBJECTIVES

Benchmarks (i.e. datasets and evaluation metrics) play a crucial role in developing and evaluating robust head pose estimation methods. A good benchmark not only allows to identify the challenges of a task but also enables the development of better methods for solving it. An in-car head pose dataset provides difficult illumination conditions, occlusions, and extreme head poses. The recent popularity of deep learning methods with their large model complexity stresses the demand for a large dataset [12]. Available head pose datasets have drawbacks in terms of size, annotation accuracy, resolution, and diversity (see Table 2.3). To close this gap, this chapter presents *DD-Pose*, a large-scale head pose benchmark.

This chapter is based on the work published in [109] (©2019 IEEE).

3.2. PROPOSED APPROACH

This chapter introduces *DD-Pose*[1], a large-scale head pose benchmark featuring driver camera images acquired during complex naturalistic driving scenarios. The proposed benchmark provides 330 k high resolution driver camera images from 27 subjects with precise continuous 6 DOF head translation and rotation annotations. *DD-Pose* includes a variety of non-frontal poses and occlusions occurring in complex driving scenarios. Occlusions from steering wheel, hands, and accessories such glasses or sunglasses are present and manually annotated as such on a per-frame basis. Sample annotations of the benchmark can be found in Figure 3.1.

[1]Available at https://dd-pose-dataset.tudelft.nl

Figure 3.1: *DD-Pose* provides precise 6 degrees of freedom (DOF) head pose annotation for 330 k stereo image pairs acquired in an in-car environment. The benchmark offers significant out-of-plane rotations and occlusions from naturalistic behavior introduced by complex driving scenarios. Annotations for partial and full occlusions are available for each high-resolution driver camera image. An additional camera capturing the interior of the car allows for further multi-sensor driver analysis tasks.

High resolution images of the driver's head are acquired by a stereo camera setup mounted behind the steering wheel. Continuous frame-wise head pose is obtained by a optical marker tracker measuring the 6 DOF pose of a marker fixated on the back of each subject's head. A per-subject transformation from the head mounted marker to the head coordinate system is found by a novel calibration device.

In addition to the driver stereo camera, the proposed setup uses a wide angle RGB camera depicting the driver from the rear side to allow for upper-body analysis of the driver action. Vehicle parameters such as steering wheel angle, velocity and yaw rate are also part of the benchmark.

All sensors are calibrated intrinsically and extrinsically, such that the coordinate transformations between their coordinate systems are known. Depth information can be extracted from the provided stereo camera images by using a disparity estimation algorithm, e.g. semi-global matching [53]. The optical marker tracker and the stereo driver camera are electrically synchronized, resulting in a head pose measurement free of drift and latency.

DD-Pose offers a broad distribution of poses and challenging lighting conditions like dark nighttime driving, tunnel entrances/exits and low standing sun. 12 driving scenarios were conducted to gain highly variant, yet naturalistic images of the driver. Nine driving scenarios comprise drives through a big German city with lane merges,

complex roundabouts, parking, and pedestrian zones with pedestrian interactions. In addition to driving scenarios, *DD-Pose* provides three standstill scenarios covering a broad range of head poses and a scenario with mobile phone use.

Overall, *DD-Pose* offers a variety of naturalistic driving data which is crucial for development and evaluation of head pose estimation algorithms in unconstrained environments. With four megapixels per camera and a mean inter-pupil distance of 274 px, *DD-Pose* offers around 60 times more face pixels than DriveAHead to extract features from fine-grained face structures such as eye gaze and evaluate whether high resolution is a benefit to the methods under test.

3.2.1. SCENARIOS

The definition of driving scenarios has an essential impact on the distribution of the head pose and textural variability of the data. E.g., a drive along the highway would be very biased towards a frontal pose and not be beneficial to train and evaluate head pose estimation methods. Non-frontal poses are favored by implicitly forcing the driver have to look out of the car, e.g. by interacting with pedestrians in a pedestrian zone, and instructing the driver to read shop names on the side of the street. Yet, to be representative of naturalistic drives, scenarios of standard traffic manoeuvres, such as passing zebra crossings, highway merges, roundabouts, parking and turning the vehicle are included. To provide more extensive poses, scenarios while standing are included, where the driver is instructed to fixate his or her gaze on predefined locations within the car, forcing large head rotations and translations, and making a phone call.

The scenarios of *DD-Pose* are defined in Table 3.1, alongside with their intended properties on data variability.

For the in-car gaze fixation scenario (Table 3.1, #9) this chapter defines the following protocol: the car stands still with the steering wheel in straight position. The subject is asked to turn the head to point at a predefined set of targets in the car. A button is to be pressed by the subject for the period he or she is fixating the object, thus annotating the time stamps of fixation ad-hoc. Among the targets are mirrors, in-car buttons and displays.

In summary, these carefully-chosen scenario definitions result in a large variance in head rotation and head translation, but also facial expressions.

3.2.2. HARDWARE SETUP AND COORDINATE SYSTEMS

A research vehicle has been equipped with a stereo camera facing the driver (each 2048x2048 px, 16 bit, IR sensitive). It is mounted near the speedometer. An infra-red LED illuminates the driver. A wide angle interior camera (RGB) captures the driver's cabin from the rear side. An optical marker tracker was mounted on the rear right behind the driver. The optical marker tracker can measure the 6 DOF pose of a marker consisting of multiple IR retroreflective spheres. The subject wears such a marker on the back of his or her head, which is fixated using a rubber band.

The driver stereo camera, LED illumination and optical marker tracker are electrically triggered at 15 Hz. The other sensors are synchronized.

A head calibration device was designed which defines the head coordinate system when attached to the driver's head while being simultaneously being measured by the

Table 3.1: Driving scenario definitions and the resulting features of the proposed benchmark. 12 scenarios are defined to implicitly enforce a broad distribution of head poses and texture.

#	Description	Rot	Trans	Occl	Stw Occl	Facial ex	Illum var	Ped inter	Remark
0	generic driving	low	low	low	med	high	med	low	talking
1	zebra crossing	low	low	low	low	med	med	high	crossings and bus stops
2	merge	high	med	low	low	med	med	low	mirrors, look over shoulder
3	tunnel	low	low	low	low	med	high	low	entrance, exit
4	roundabout	high	med	low	med	med	med	low	also multi-lane roundabout
5	ped zone	high	med	low	high	med	med	high	incl. two-step turn
6	intentional occl	med	med	high	med	high	med	low	occlusions, facial expressions
7	shop name reading	high	med	med	low	high	med	high	shops left and right
8	parking	high	high	med	high	high	med	med	parking in
9	in-car fixation	high	med	med	no	high	med	low	no driving
10	large translations	med	high	med	no	med	med	low	no driving
11	large rotations	high	med	med	no	med	med	low	no driving
12	hand-held calling	high	med	high	no	med	med	low	no driving

Rot: rotation; Trans: translation; Occl: occlusion; Stw occl: steering wheel occlusions;
Facial ex: facial expressions; Illum var: illumination variance; Ped inter: pedestrian interaction.

optical marker tracker.

Each camera, the optical marker tracker, the head mounted marker, the driver's head and the car's chassis define a coordinate system. A transformation between two coordinate systems A and B is defined as a homogeneous matrix $T^{\mathrm{A}\leftarrow\mathrm{B}}$ which transforms a homogeneous point p^{B} into p^{A} by $p^{\mathrm{A}} = T^{\mathrm{A}\leftarrow\mathrm{B}} \cdot p^{\mathrm{B}}$.

See Figure 3.2 for a visual overview of the sensors, their coordinate systems and the transformations in between them.

3.2.3. OPTICAL MARKER TRACKER TO DRIVER CAMERA CALIBRATION

$T^{\mathrm{cam\_driver\_left}\leftarrow\mathrm{marker\_tracker}}$ and the camera intrinsic parameters are obtained simultaneously by a calibration routine which makes use of 3D checkerboard corner positions. The 3D checkerboard corner positions inside the marker tracker coordinate system are defined by attaching retro-reflective spheres to the checkerboard, thus making it a marker measurable by the optical marker tracker. With the 3D checkerboard corner positions and their corresponding 2D projections in the image, a bundle adjustment method is used to optimize intrinsic and extrinsic camera parameters, such as focal lengths, principal points, distortion parameters and rectification parameters [143]. $T^{\mathrm{cam\_driver\_left}\leftarrow\mathrm{marker\_tracker}}$ is obtained as a by-product of the optimization.

3.2.4. MARKER TO HEAD CALIBRATION

The head coordinate system is defined as follows. The origin is located in the nasion of the head. The x-axis points in frontal direction. The y-axis points towards the left ear. The z-axis points upwards; it touches the chin centrally. The xz-plane mirrors the head.

A calibrator is designed to attach to the driver's head during the per-subject calibration process. It provides a notch to touch the nasion. A chin slider is adjusted such that it touches the chin centrally. Two cheek sliders are slid against the head such that they touch the cheeks with equal force, thus defining symmetry about the xz-plane. It is also

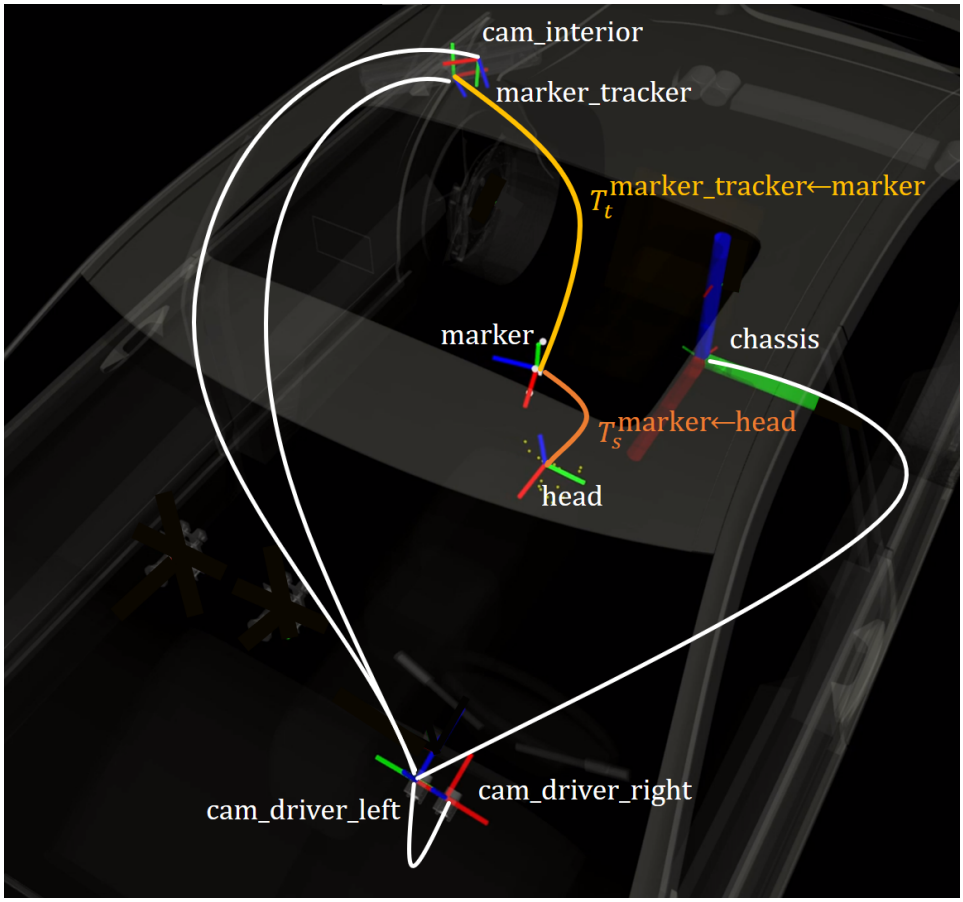


Figure 3.2: In-car hardware setup, coordinate systems and transformations. White arcs denote static transformations acquired once during the setup calibration process. The yellow arc denotes the transformation $T_t^{\text{marker\_tracker}\leftarrow\text{marker}}$ being measured by the optical marker tracker for each frame at time t. The orange arc denotes the transformation $T_s^{\text{marker}\leftarrow\text{head}}$ being calibrated once per subject s. All transformations are provided with *DD-Pose*.

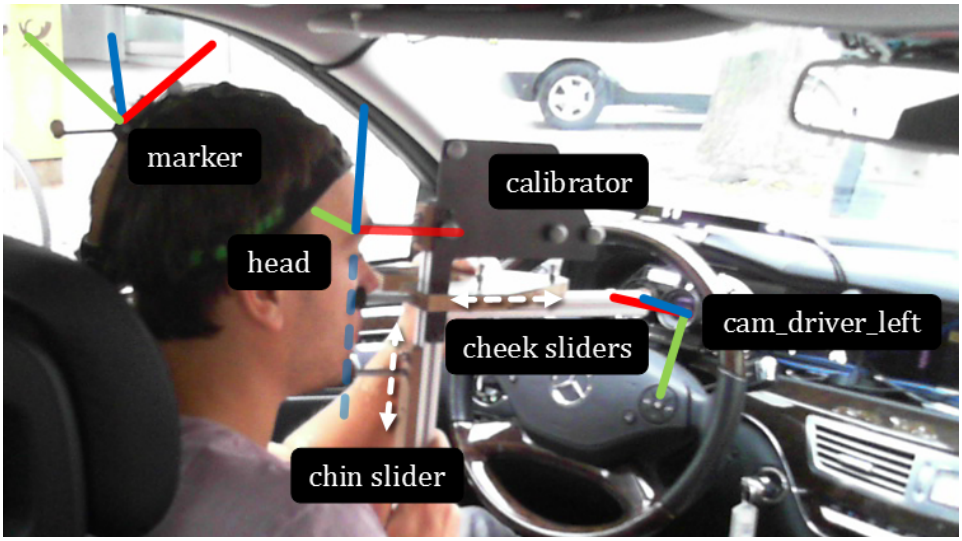


Figure 3.3: The per-subject head calibration process. A calibrator whose pose can be measured by the optical marker tracker is attached to the head by touching the nasion and both chin and cheek sliders in proper position.

equipped with retroreflective spheres such that its pose can be measured by the optical marker tracker. Its coordinate system is defined such that it coincides with the head coordinate system above. When it is attached properly, the per-subject transformation between marker and head is then $T_s^{\text{marker}\leftarrow\text{head}} := T_t^{\text{marker}\leftarrow\text{calibrator}}$. This process has to be performed once per subject and is valid as long as the marker is fixated at the subject's head. The calibration process is illustrated in Figure 3.3.

3.2.5. DATA PREPROCESSING

Depending on the driver's head pose, the retroreflective spheres of the head-worn marker are visible in the camera image. They are removed to avoid models to overfit on these. This chapter extends the approach of [119], where the projected locations of the spheres are filled with interpolations of the values of their surroundings. As markers will mostly be hidden behind the subjects' head, a heuristic is employed to only blur the spheres which are likely visible. The heuristic is based on an empirically found range of head poses and conservatively set, i.e. rather fill hair or face border than leave spheres visible.

3.2.6. OCCLUSION ANNOTATIONS

Each driver camera image is manually annotated for its occlusions based on the visibility of facial landmarks, as defined in [115]. **none**: all 68 landmarks visible; **partial**: at least one landmark occluded; **full**: all landmarks occluded.

3.2.7. DATASET SPLITS

To allow for a fine-grained evaluation, this chapter splits the data into the disjoint subsets *easy*, *moderate*, and *hard* depending on the angular distance of the measured head pose from a frontal pose (looking directly into the driver camera) α_f and the presence of occlusion. **easy**: $\alpha_f \in [0,35)^\circ \wedge \text{occl} \in \{\text{none}\}$; **moderate**: $(\alpha_f \in [0,35)^\circ \wedge \text{occl} \in \{\text{partial}\}) \vee (\alpha_f \in [35,60)^\circ \wedge \text{occl} \in \{\text{none, partial}\})$; **hard**: $\alpha_f \in [60,\infty)^\circ \vee \text{occl} \in \{\text{full}\}$;

3.3. DATASET ANALYSIS

DD-Pose comprises recordings of 27 subjects, of which 21 are male and 6 are female. The average age is 36 years. The youngest and oldest driver are 20 and 64 years old.

There are 330 k measurements of the driver stereo image camera along with interior camera images. Head pose measurements are available for 93% of the images. The proportion of the dataset splits is (*easy*, *moderate*, *hard*) = (55%, 33%, 12%).

For the left driver camera images, 5% are fully occluded, 19% are partially occluded (not counting glasses or sun glasses) and 76% have no occlusion. In 41% of the images, the driver wears glasses, in 1% sunglasses.

There are 13 scenarios, out of which nine are driving scenarios (#0 - #8) and four are non-driving scenarios (#9 - #12); see Table 3.1. The shortest scenario (#3, tunnel entrance/exit) is on average 24 s long. The longest scenario (#5, pedestrian zone) is on average 211 s long.

The mean inter-pupil distance is 274 px (cf. DriveAHead: 35 px [119]).

The distribution of head rotation angles of *DD-Pose* and DriveAHead [119] is depicted in Figure 3.4. The angles vary in the following ranges, ignoring outliers with less than 10 measurements in a 3° neighborhood: roll $\in [-63..60]^\circ$; pitch $\in [-69..57]^\circ$; yaw $\in [-138..126]^\circ$. The mean pitch angle is -20°, caused by the driver camera mounted at the speedometer.

The distribution of head translation occurrences of *DD-Pose* and DriveAHead [119] is depicted in Figure 3.5. *DD-Pose* covers a broad volume of head locations.

Overall, *DD-Pose* offers an order of magnitude more data for off-centered head poses than comparable datasets [119].

3.4. EXPERIMENTS

To show that the proposed benchmark contains challenging imagery and head poses, this chapter evaluates the performance of two 6 DOF off-the-shelf head pose estimation methods on it. In addition to the experiments conducted in this chapter, a novel method for head pose estimation which employs the dataset will be presented in the next chapter.

3.4.1. OFF-THE-SHELF HEAD POSE ESTIMATION METHODS

One method is the head pose prior, which always assumes the head to be present in the mean pose obtained from the dataset. The second method performs head pose estimation by localizing facial landmarks and solving the Perspective-n-Point (PnP) problem.

Prior: on a dataset with a large amount of frontal poses, this method is expected to perform very well, despite performing bad on rare poses. The mean head translation

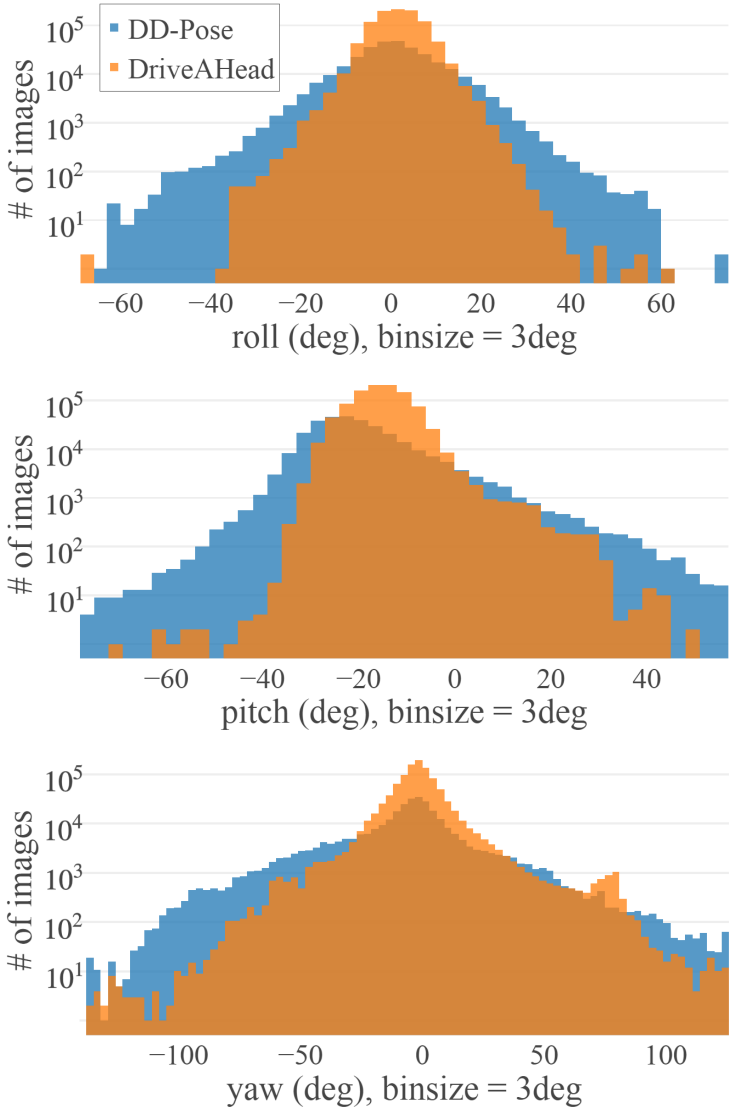


Figure 3.4: Distribution of head rotation angles of the proposed benchmark *DD-Pose* and DriveAHead [119] with respect to a frontal pose into the camera. While both datasets cover a broad range of rotations, *DD-Pose* supplies an order of magnitude more data for non-frontal head rotations.

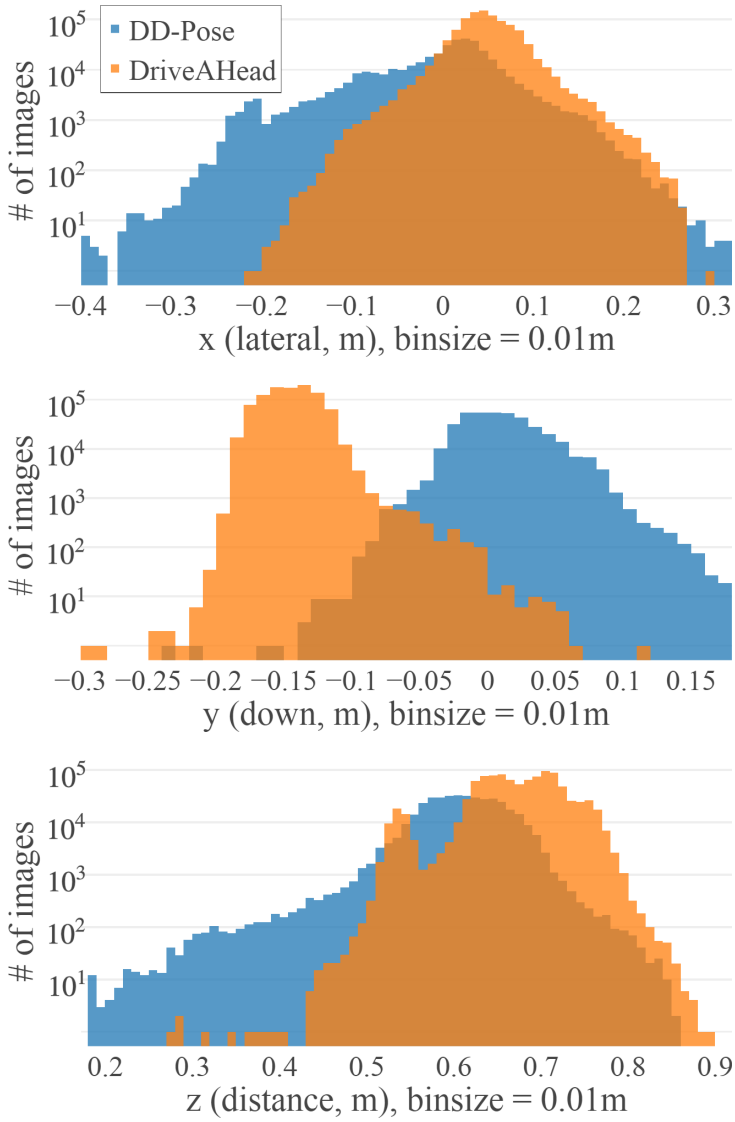


Figure 3.5: Distribution of head translations of *DD-Pose* and DriveAHead [119] in the camera coordinate system. Although the action volume of the driver is limited in the driver's seat, the datasets differ in their translation distribution. *DD-Pose* covers a larger lateral space, is unbiased in y-direction and also depicts very close-by heads.

of *DD-Pose* wrt. the camera is $\bar{t} = (0.011\,\text{m}, 0.006\,\text{m}, 0.608\,\text{m})$. The mean rotation is $yaw = -6.6°$, $pitch = -20.1°$, $roll = 0.7°$.

OpenFace 2.0: the second method this chapter evaluates is OpenFace 2.0 [7], a State-of-Art (SoA) face analysis toolkit. Head pose estimation is performed by localization of facial landmarks via Convolutional Experts Constrained Local Model (CE-CLM). The facial landmarks are assigned to a 3D landmark model in head coordinates. The pose is found via solving the Perspective-n-Point (PnP) problem, i.e. finding the pose of the head coordinate system with respect to the camera coordinate system which minimizes the projection error. Pretrained models from the authors [7] are used, and the pose is transformed such that it fits the head coordinate system defined above. The model uses multi-view initialization to account for extreme poses.

3.4.2. EVALUATION METRICS

Evaluation metrics play an important role on evaluating the performance of the methods for the specific task. The task of head pose estimation is evaluated for translation and rotation separately.

Recall: recall defines on which percentage of the images a head hypothesis from head pose estimation method exists. Images without a hypothesis are left out when evaluating translation and rotation.

Translation: the mean Euclidean distance for each axis and the mean absolute error (translation) (MAE_t), i.e. mean Euclidean distance, between ground truth head origin and hypothesis head origin.

Rotation: the commonly used metric mean absolute error (rotation) (MAE_R) can be performed on each of the three rotation angles separately or by computing a single rotation angle between ground truth and hypotheses (geodesic distance). In both cases, outliers will have a small weight on biased datasets, e.g. with many frontal poses and a few extreme poses. For an unbiased evaluation of head rotation, balanced mean absolute error (BMAE) is used, as introduced in [119]. It splits the dataset in bins based on the angular difference from the frontal pose and averages the MAE_R of each of the bins:

$$\text{BMAE}_{d,k} := \frac{d}{k} \sum_i \phi_{i,i+d} \; \forall i \in d\mathbb{N} \cap [0, k] \tag{3.1}$$

where $\phi_{i,i+d}$ is the MAE_R of all hypotheses, where the angular difference between ground truth and frontal pose is between i and $i+d$. During evaluation, this chapter uses bin size $d := 5°$ and maximum angle $k := 75°$.

3.4.3. RECALL

The prior method, by construction, has a recall of 1.0. The recall of OpenFace 2.0 on the whole dataset is 0.76 and for the subsets (*easy, moderate, hard*) = (0.95, 0.65, 0.16). A more fine grained analysis on the recall value depending on the angular distance from the frontal pose is found in Figure 3.6. One can see the influence of the definition of the subsets. While the *easy* subset offers a large recall as it covers unoccluded heads with angles up to 35°, the *moderate* subset covers the partial occlusions in this range with a lower recall. The overall recall drops with increasing angle.

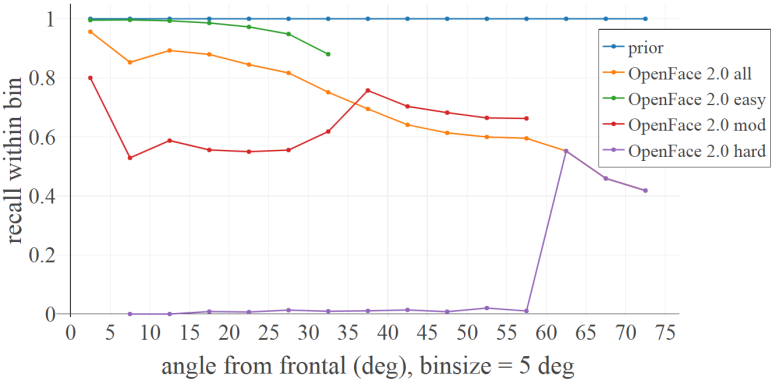


Figure 3.6: Recall depending on angular difference from frontal pose. The recall of OpenFace 2.0 on the whole dataset drops with increasing rotation from the frontal pose.

Table 3.2: Translation errors (mm). Errors along all axes and mean Euclidean Distance MAE_t for the subsets.

Subset	Prior				OpenFace 2.0			
	x	y	z	MAE_t	x	y	z	MAE_t
all	40	21	36	66	8	8	41	44
easy	23	19	32	49	5	6	31	33
moderate	54	21	38	78	12	10	58	63
hard	83	27	46	107	44	30	134	148

3.4.4. TRANSLATION ERROR

The errors in head translation estimation are listed in Table 3.2. The errors on the prior method implicitly denote statistics of the distribution of the subsets. The MAE_t increases from 5 cm to 11 cm from the easy to the hard subset, caused by a larger translation variance around the mean translation in the measurements. OpenFace 2.0 localizes the head translation in x and y direction for the *easy* and *moderate* subsets within 1 cm, increasing up to 4 cm for the *hard* subset. OpenFace 2.0 has approximately 4-5 times larger errors in z direction than for the other two dimensions.

3.4.5. ROTATION ERROR

An overview of the MAE_R and BMAE of the methods on *DD-Pose* is given in Table 3.3. Figure 3.7 depicts the MAE_R depending on the angular difference from a frontal pose.

The prior method implicitly denotes statistics on the rotation measurement distribution around the mean rotation. The MAE_R increases from 11° to 45° from the easy subset to the hard subset, showing the increasing variance for the more difficult subset.

The MAE_R of OpenFace 2.0 ranges from 5° on the *easy* subset to 33° on the *hard* subset, i.e. the error increases by more than a factor of 6 when facing more challenging poses and occlusions. For comparison: the reported MAE_R of OpenFace 2.0 is 2.6° on the BU

Table 3.3: Overall mean absolute error (rotation) (MAE_R) and balanced mean absolute error (BMAE) in degrees of the head pose estimation methods for the subsets; MAE_R for roll, pitch, yaw of OpenFace 2.0 (deg).

Subset	Prior		OpenFace 2.0				
	MAE_R	BMAE	MAE_R	BMAE	roll	pitch	yaw
all	20	32	9	16	5	4	4
easy	11	14	5	5	3	3	2
moderate	27	26	14	13	8	6	8
hard	45	34	33	31	13	9	27

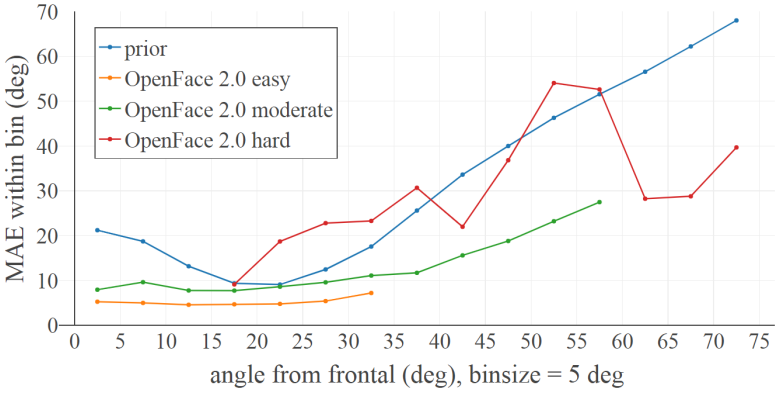


Figure 3.7: Mean absolute error (rotation) (MAE_R). All methods increase in terms of MAE_R for more extreme poses.

dataset [64] and 3.2° on the ICT-3DHP dataset [7].

The evaluations show, that the data of the proposed benchmark provide challenges to head pose estimation methods due to its broad distribution of angles.

3.5. ADOPTION OF *DD-Pose* SINCE RELEASE

DD-Pose has been adopted by the scientific community since the release on the public website[2] in October 2019. See Figure 3.8 for a screenshot of the public website. As of May 2023, 58 persons have registered for the use of the dataset. The registrants come from 23 different countries, most prominently from China.

Sign-up is restricted for scientific, non-commercial purposes only, as stated by the license terms on the website[3]. The license is derived from the EuroCity Persons Dataset license[4] [12] and is strict to ensure compliance with general data protection regulation (GDPR) of the European Union. This led to the denial of 64 interested parties who have

[2] https://dd-pose-dataset.tudelft.nl
[3] https://dd-pose-dataset.tudelft.nl/eval/license/license
[4] https://eurocity-dataset.tudelft.nl/eval/license/ecplicense

not fulfilled the license requirements.

Overall, an international audience has shown interest in the dataset, considering the size of the research field of camera-based head pose estimation.

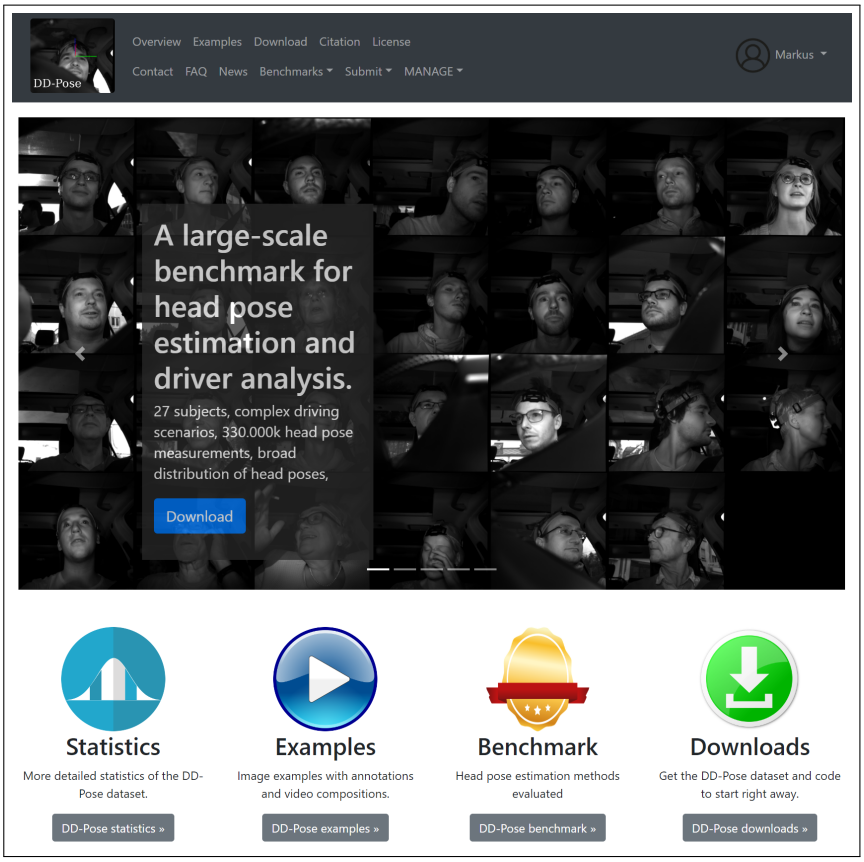


Figure 3.8: Screenshot of the public website of *DD-Pose* (https://dd-pose-dataset.tudelft.nl). Visitors get an overview of dataset statistics and can download the dataset upon successful registration.

4

MONOCULAR DRIVER 6 DOF HEAD POSE ESTIMATION LEVERAGING CAMERA INTRINSICS

This chapter proposes a new method for camera-based head pose estimation. It leverages the dataset presented in Chapter 3. Head pose of a driver will be used as a cue for the driver's awareness of a pedestrian in Chapter 6.

4.1. OBJECTIVES

Head pose estimation plays an essential role in human understanding, as it is our natural cue for inferring focus of attention, awareness, and intention. For machine vision, the task is to estimate both the translation and rotation of the head from camera images.

Head pose constitutes up to 3 degrees of freedom (DOF) for translation and up to 3 DOF for rotation. Unlike most previous work that has estimated only a subset of the 6 DOF, e.g., by not estimating translation, estimating less than 3 DOF of rotation, or by estimating coarse bins of rotation, this chapter focuses on *full* 6 DOF on a *continuous* scale, as required by most of the applications mentioned in Chapter 1. Both the estimation of translation and rotation can be seen as a regression problem, based on the intensity input image.

Estimating a metric translation (i.e., $[x, y, z]$ in meters) from monocular images requires (implicit) knowledge of the head size because head size and distance (as a part of translation) are inversely proportional with respect to pinhole camera projection: a larger head will appear closer. Parameters determining how the head is being projected into the camera image are the intrinsic camera parameters, i.e., focal lengths and principal point. These factors directly affect the estimation accuracy. This chapter will show that a head pose estimation method needs to be *intrinsics-aware* for precise estimation, i.e., being a *camera*-based method rather than an intrinsics-agnostic *image*-based method. On the contrary, a head pose estimation method which does not explicitly consume camera pa-

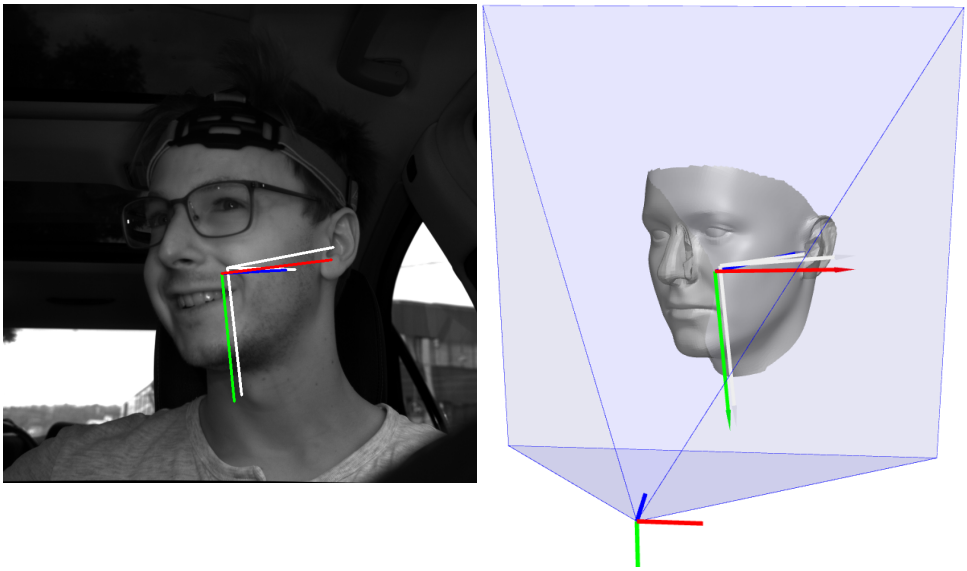


Figure 4.1: intrApose is an **intr**inics-**A**ware head **pose** estimation method. The method estimates continuous 6 degrees of freedom (DOF) head pose (rotation and translation) from a single intensity image and known camera intrinsics. Left: Input intensity image (*DD-Pose* validation set). Right: 3D scene with camera frustum (blue). Face mesh and RGB axes: 6 DOF head pose result of intrApose. Gray axes: ground truth head pose.

rameters encodes implicit assumptions which hinders generalization to different camera setups.

There are different representations for 3 DOF rotations, most commonly Euler angles, and Quaternions. Both come with discontinuities (i.e., Euler angles between 359° and 0°) and non-linearities. Rotation estimation methods have applied workarounds for dealing with the discontinuity of rotation representations rather than intrinsically using a continuous representation, that is more suitable for deep learning methods [16, 145].

This chapter presents intrApose, a method for full, continuous 6 DOF head pose estimation which explicitly leverages camera intrinsics and uses a continuous rotation representation. It operates directly on intensity images and camera intrinsics without a previous face detection or landmark estimation step.

Training and evaluating a method that estimates full, continuous 6 DOF head pose demands a dataset that provides camera intrinsics alongside continuous 3 DOF translation annotation and continuous 3 DOF rotation annotation. To that end, this chapter bases the experiments on the driver head pose dataset *DD-Pose* which was introduced in Chapter 3. See Figure 4.1 for exemplary input to and output of the presented method.

This chapter is based on the work published in [108] (©2023 IEEE).

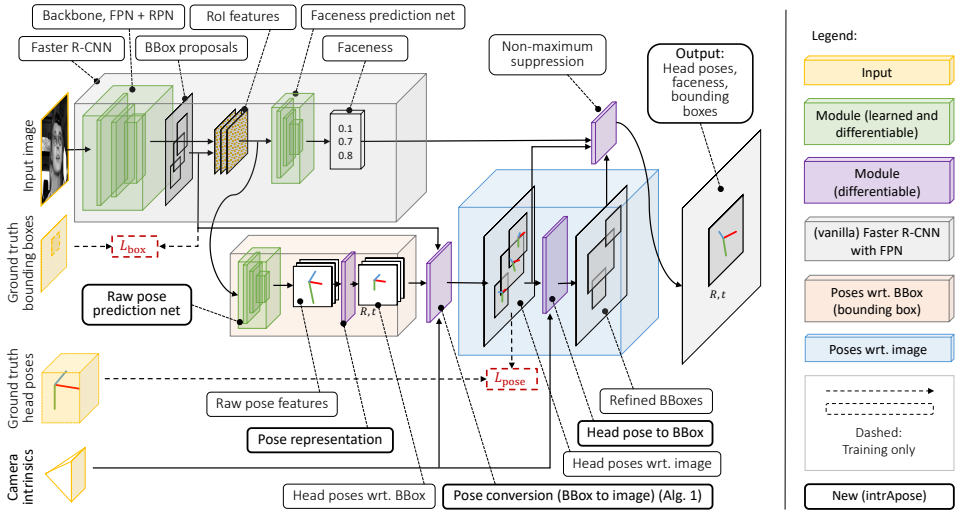


Figure 4.2: Architectural overview of intrApose, the proposed head pose estimation method, with novel parts highlighted in bold. intrApose takes intensity images and camera intrinsics as an input. RoI features are obtained from BBox (bounding box) proposals based on Faster R-CNN with feature pyramid network (FPN) [73, 105] (gray box). A **Raw pose prediction net** regresses raw pose features which the **Pose representation** module converts to head poses $\in \mathrm{SE}(3)$, i.e., the rotation R spans an orthonormal basis. Up to here, the poses are relative to their respective BBox (orange box). The BBox-local poses are converted to be image-global (blue box); see Algorithm 1 and Figure 4.4. **Bounding boxes are obtained based on the predicted head poses**. A non-maximum suppression step filters overlapping predictions. The output is a set of head poses, faceness scores and bounding boxes. During training, losses are applied to BBox proposals and head poses (dashed lines). The whole architecture is *intrinsics-aware*, specifically in the **Pose conversion (BBox to image)** module and the **Head pose to BBox** projection module, but also with respect to augmentations (see Section 4.2.5) and cropping/resizing.

4.2. PROPOSED APPROACH

4.2.1. OVERVIEW

This chapter proposes intrApose, a novel method for image-based driver head pose estimation based on a deep neural network that regresses continuous 6 DOF from a single intensity image without prior face detection or landmark estimation (see Figure 4.2). The main building blocks are a Faster R-CNN-based network which regresses BBoxes and extracts RoI features within these. intrApose learns raw pose features and converts them to a continuous, full 6 DOF head pose within the BBox. This BBox-local pose is converted to an image-global pose in the camera frame while respecting camera intrinsics (see Algorithm 1). Using differentiable modules and a continuous rotation representation allow for a plain overall architecture.

The proposed method is inspired by the recent head pose estimation method img2pose proposed by Albiero *et al.* [4]. The latter presents an efficient Faster R-CNN-based model which regresses 6 DOF head poses without prior face detection or landmark localization. The method has shown strong performance on datasets with ground truth head poses obtained from manually annotated facial landmarks.

The main differences are: (a) intrApose is camera-intrinsic aware: focal lengths are consistently used as opposed to using image size as a heuristic for focal length. (b) intrA-

pose uses a continuous rotation representation which makes both pose normalization and usage of a calibration point loss as employed by img2pose obsolete, therefore simplifying the architecture. (c) intrApose provides an architecture with a differentiable pose conversion which makes an inverse image-to-bbox pose conversion step (i.e., inverse of Algorithm 1) at training time superfluous, therefore further reducing model implementation complexity. (d) intrApose uses Faster R-CNN anchor box aspect ratios and sizes tuned for human heads, as opposed to aspect ratios and sizes of generic objects, such as cars or cats.

4.2.2. DEFINITION OF HEAD POSE

In this thesis, *head pose* is defined as a linear transformation matrix $T^{\text{cam}\leftarrow\text{head}} \in \text{SE}(3)$, the special Euclidean group, which transforms a three-dimensional homogeneous point $p^{\text{head}} = \left[x^{\text{head}}, y^{\text{head}}, z^{\text{head}}, 1\right]^T$ given in the *head* coordinate frame to a point $p^{\text{cam}} = [x^{\text{cam}}, y^{\text{cam}}, z^{\text{cam}}, 1]^T$ in the *cam* coordinate frame by $p^{\text{cam}} = T^{\text{cam}\leftarrow\text{head}} \cdot p^{\text{head}}$, thus representing translation by 3 DOF and rotation by 3 DOF on a continuous scale.

Transforms $T \in \text{SE}(3)$ are constructed as in (4.1). They can be decomposed into a 3×3 submatrix $R \in \text{SO}(3)$ representing the rotation and a translation vector $t = [t_x, t_y, t_z]$. Ultimately, a homogeneous point multiplied from the right-hand side will be rotated by R and afterward shifted by t.

$$T^{\text{cam}\leftarrow\text{head}} = \begin{bmatrix} & & & t_x \\ & R & & t_y \\ & & & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.1}$$

For the *cam* frame, this thesis follows the convention: x to the right, y to the bottom, and z in the viewing direction. The *head* frame can be an arbitrary, head static frame.

One common convention is $head^C$ (C as in camera-like): x sinister, y inferior, z posterior, with the origin being close to the nasal point (though there is no agreed-upon convention for the origin). This definition has the advantage, that a head pointing straight towards the camera will have a null rotation (identity). However, it has the drawback, that the Tait-Bryan rotation components roll (cervical side-bending), pitch (cervical flexion), and yaw (cervical rotation) do not correspond to the axes x, y, and z, respectively.

4.2.3. WHY CAMERA INTRINSICS ARE ESSENTIAL FOR POSE ESTIMATION

The camera intrinsic matrix K defines, how 3D points in the *cam* frame are projected onto a rectified image. See (4.2):

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4.2}$$

It is a 3×3 matrix consisting of focal lengths f_x and f_y for x and y axes and principal point (c_x, c_y) representing the optical center within the image. The axis skew parameter s is typically assumed 0. K projects a point $p^{\text{cam}} = [x^{\text{cam}}, y^{\text{cam}}, z^{\text{cam}}]^T$ given in the *cam* frame onto pixel coordinates $[u, v]$ by $[u \cdot w, v \cdot w, w]^T = K \cdot p^{\text{cam}}$. Points residing in a

different frame, e.g., *head*, can be transformed into the *cam* frame by a rigid transform $T^{\text{cam}\leftarrow\text{head}} \in \text{SE}(3)$. K is non-singular: its inverse K^{-1} projects an image coordinate $[u, v, 1]$ into a 3D ray representing all points in *cam* frame which project onto $[u, v, 1]$.

Translation and Rotation Errors: One implication of *assuming* focal lengths not corresponding to the camera optics, e.g., kf_x and kf_y for a factor $k \in \mathbb{R}$ will project a central object to k times the image size (k^{-1} times the distance to the camera), compared to focal lengths f_x and f_y. When estimating object pose, its z translation will be k times as large.

Another implication is that assuming a wrong camera intrinsic matrix affects rotation estimations which are more apparent at the image border. Take the example in Figure 4.3: two cameras differing in focal lengths by a factor of two are positioned such that their projections of a head pose into the respective camera images are approximately equal (close to the right image border). The camera with a larger field of view (smaller focal length) is closer to the head and rotated. A pose-from-image estimation using these different camera intrinsics from the same image results in a translation error of factor two and a rotation error of $> 11°$ (in this example).[1]

4.2.4. PROPOSED MODEL

See Figure 4.2 for an architectural overview. Given an image $\mathbf{I}$, intrApose estimates full 6 DOF continuous head pose $T_i^{\text{cam}\leftarrow\text{head}}$ for each head i within the image $\mathbf{I}$. The major building blocks are a Faster R-CNN module which predicts bounding box proposals along with a faceness score. The prediction head performs RoI pooling on the backbone's feature maps based on the bounding box proposals to obtain RoI features. A *raw pose prediction net* predicts an intermediate, unconstrained representation of *raw pose features* which are typically of small size (such as six to 12 values, see Table 2.1) representing rotation and translation. The *Pose representation* converts the potentially degenerate raw pose features to a head pose $\in \text{SE}(3)$ wrt. the isolated BBox. A *Pose conversion* module converts BBox-local head pose to an image-global pose such that it projects approximately equal within the (whole) image. This essentially performs scaling and rotation of the pose based on image intrinsics and bounding box location and size (see Algorithm 1 and Figure 4.4). The final step is a non-maximum suppression based on projected bounding boxes and faceness scores and yields a set of head poses. Let us describe the components in more detail.

Backbone, FPN + RPN: intrApose extends the two-stage object detection approach of Faster R-CNN [105] with Feature Pyramid Networks (FPN) [73] by a head pose estimation module. Faster R-CNN consists of a backbone network that extracts features on multiple scales from the input image. Using these features and anchor bounding boxes of typical object aspect ratio and shape, a region proposal network (RPN) predicts bounding box proposals alongside an objectness score. An RoI pooling operation aggregates features for each bounding box proposal into *RoI features*.

Head pose estimation module: As in img2pose [4], this chapter proposes a network that regresses a BBox-local head pose and a faceness score p_i for each RoI feature map i. In contrast to img2pose, which regresses 6-element vectors representing head pose directly (rotvec and translation), the proposed architecture allows for a generic scheme

[1] The estimated poses using camera A and camera B would differ by the rigid transform which brings camera A into camera B.

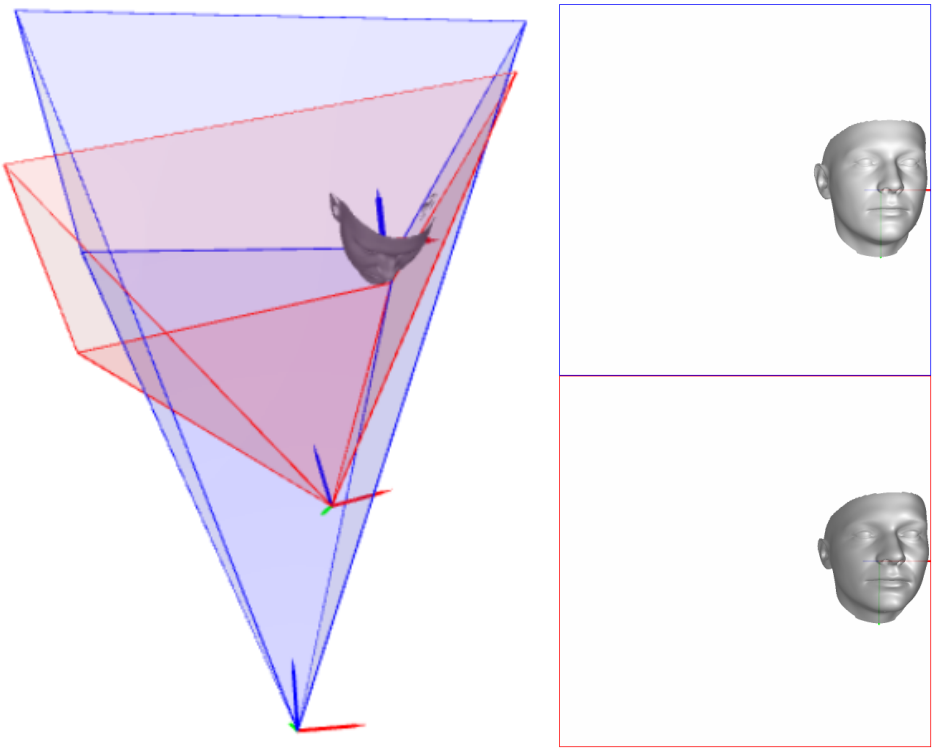


Figure 4.3: Motivational example of translation and rotation error introduced by assuming different focal lengths resulting in similar projections. Left: Frusta of two cameras resulting in approximately the same projection of face on the right border of the image. Blue: frustum of camera A with focal length f. Red: frustum of camera B with focal length $f/2$. Camera B is closer to the object due to the larger field of view. In this example, it is rotated against camera A by $> 11°$. The right half depicts the projections into the image space, with camera A on top and camera B with a larger field of view on the bottom.

by estimating an intermediate raw pose feature representation by a *Raw pose prediction net* that is being converted to a head pose $T_i^{\text{cam}\leftarrow\text{head}} \in \text{SE}(3)$ by a differentiable *Pose Representation* module.[2]

Raw pose prediction net: The raw pose prediction net estimates unconstrained, raw pose features $f_{\text{R,t}}$ (R: rotation, t: translation) wrt. the BBox for each RoI feature map. It consists of a batch-normalized fully connected layer with 256 features and ReLU activation followed by another fully connected layer reducing to the number of raw pose features $f_{\text{R,t}}$.

Pose representation: The *Pose representation* module converts the raw pose features into a $T_{\text{bbox}}^{\text{cam}\leftarrow\text{head}} \in \text{SE}(3)$ representation. Section 2.1.1 showed that there is a number of pose representations available. As Zhou *et al.* [144] and Levinson *et al.* [69] pointed

[2] In the case of img2pose [4], *raw proposal pose features* are a 6-vector $(r_x, r_y, r_z, t_x, t_y, t_z)$ with (r_x, r_y, r_z) being a rotation vector (rotvec). The *Pose Representation* module converts the rotation vector to a rotation matrix by standard means (Rodrigues' rotation formula).

out, a continuous, differentiable rotation representation is favorable. This chapter will analyze different rotation representations, such as the (discontinuous) rotation vector representation of Albiero *et al.* [4], but also the Gram-Schmidt-based rotation representation Ortho6D proposed by Zhou *et al.* [144], and the symmetric-orthogonalization based rotation representation SVDO$^+$ proposed by Levinson *et al.* [69]. The translation part of the pose is treated in a regular manner, meaning 3 DOF metric translation is being regressed.

From an integration perspective, both Ortho6D and SVDO$^+$ take six, respectively nine unconstrained values as an input and create a rotation matrix $R \in \mathrm{SO}(3)$. One important aspect is that the pose representation needs to be differentiable to allow for gradients to pass during training. Both Ortho6D and SVDO$^+$ are differentiable.

A common practice when predicting rotation is to estimate the delta from a normalized rotation (zero mean, unit standard deviation). Normalization would in principle happen within this module, though experiments showed that it is unnecessary with a continuous rotation representation. Note that using a continuous pose representation allows designing a network without bells and whistles, i.e. no coarse-to-fine approach, no estimation of delta to a mean pose, no dealing with values at discontinuities, etc.

Pose conversion (BBox to image): As posed by Albiero *et al.* [4], each BBox proposal is a cut-out of the image and lost its information about the location within the image. Therefore, head poses are estimated wrt. their BBox and need conversion to the full image. To that end, this chapter propose an intrinsics-, crop- and scale-aware BBox pose to image pose conversion method that extends the conversion method of img2pose [4] and is described in Algorithm 1 and illustrated in Figure 4.4. In essence, the conversion method builds a homogeneous canonical BBox camera matrix K_{bbox} which has the same focal length as the image camera matrix K_{image}, and the principal point in the bounding box center. The distance-to-camera t_z is being scaled by the ratio of image focal length to bounding box size. Therefore, BBox head pose is estimated within a canonical BBox camera. Scaling accounts for the fact that a cut-out with a close-by head is tightly enclosed by the bounding box and reflects a head further away in the image. Inter-individual head sizes are learned implicitly from the training data. The pose is afterward projected into the pixel space with K_{bbox} and back into the 3D space of the camera with the inverse of K_{image}. The homography $K_{\text{image}}^{-1} K_{\text{bbox}}$ is not orthonormal, meaning it does not keep the basis vectors of the transform orthogonal and in unit length. Therefore, a successive (differentiable) orthogonalization of the rotation is necessary to stay within SE(3).[3] Overall, Algorithm 1 makes the method intrinsics-, crop- and scale-aware.

Head Pose to Box: With head pose and camera intrinsics, well-defined bounding boxes can be obtained at minimum additional cost. If a bounding box is defined as a rectangle in the image which encloses all parts of the object of interest, then 3D points can be defined representing extrema in the head frame (chin to the top of forehead, nose, ears), and transformed into the camera frame and projected into image space using K_{image}. The image bounding box is defined by the extrema of the projected pixel coordinates. A margin can be defined either in 3D space (by taking 3D points outside a typical head), or

[3] Note that Albiero *et al.* [4] do not explicitly formalize the orthogonalization of the degenerate rotation. In their reference implementation, orthogonalization happens implicitly during the pose conversion step from rotation matrix to rotation vector (`rot_mat_to_rot_vec()`). This thesis explicitly formalizes this step.

Algorithm 1 Pose conversion (BBox to image).

```
def convert_pose_bbox_to_image(T_cam_head, bbox, K_image):
  # create bbox intrinsic matrix with same focal lengths
  # as image and principal point in center of bbox
  K_bbox = copy(K_image)
  K_bbox(cx) = get_center_u(bbox)
  K_bbox(cy) = get_center_v(bbox)

  # scale: ratio of image focal length and
  # bbox size (canonical bbox camera)
  f_image = K_image(fx)
  size_bbox = get_w(bbox) + get_h(bbox)
  scale = f_image / size_bbox
  T_cam_head(z) *= scale  # scale z

  # apply 4x4 homography matrix to head pose
  H_image_bbox = homogen(inv(K_image) @ K_bbox)
  T_cam_head = H_image_bbox @ T_cam_head
  T_cam_head = orthogonalize_svdo+(T_cam_head)

  return T_cam_head  # image-global
```


in image space (by adding a margin to the projected bounding box).

Formally: N typical extrema points $ps^{\text{head}}_{\text{extrema}} \in R^{N\times[xyz1]}$ given in the *head* frame, are transformed to the *cam* frame by the head pose $T^{\text{cam}\leftarrow\text{head}}$: $ps^{\text{cam}}_{\text{extrema}}{}^{T} = T^{\text{cam}\leftarrow\text{head}} \cdot ps^{\text{head}}_{\text{extrema}}{}^{T}$. These can be projected into the camera image with the camera matrix K_{image} by $[us, vs, 1s]^T \sim K_{\text{image}} \cdot ps^{\text{cam}}_{\text{extrema}}{}^{T}$. The bounding box B is then [min(us), min(vs), max(us), max(vs)], following the [left, top, right, bottom] convention.

Training objective: During training, the following objectives are optimized: (a) RPN bounding box proposals (vanilla Faster R-CNN), (b) RPN objectness score (vanilla Faster R-CNN), (c) Faceness score, and (d) Head pose outputs wrt. image, which is a multi-task problem. For (a) and (b) one can refer to Ren *et al.* [105] (L_{bbox}: smooth L_1 on positive samples; $L_{\text{objectness}}$: binary cross entropy L_{cls}). For the other representations, this chapter defines the following loss functions.

Faceness score: ground truth bounding boxes (automatically generated from head pose ground truth) are matched with proposal bounding boxes using Intersection over Union (IoU). Positive matches yield a faceness loss of $L_{\text{face}} = L_{\text{cls}}(p_i, 1)$ for the predicted faceness score p_i. Negative matches get $L_{\text{face}} = L_{\text{cls}}(p_i, 0)$.

Head pose: The head pose matrix $T^{\text{cam}\leftarrow\text{head}}$ can be decomposed into a 3×3 rotation matrix R and a translation vector $t = [t_x, t_y, t_z]^T$ as in (4.1). Positive matches are considered for a head pose loss $L_{\text{pose}} = L_{\text{R}} + L_{\text{t}}$, consisting of rotation loss L_{R} and translation loss L_{t}. The loss is applied to predicted poses wrt. full image.[4]

[4]This is a major difference to img2pose, which applies the loss in BBox domain, and needs an additional image-to-bbox pose conversion during training, that in-turn increases the complexity of the architecture.

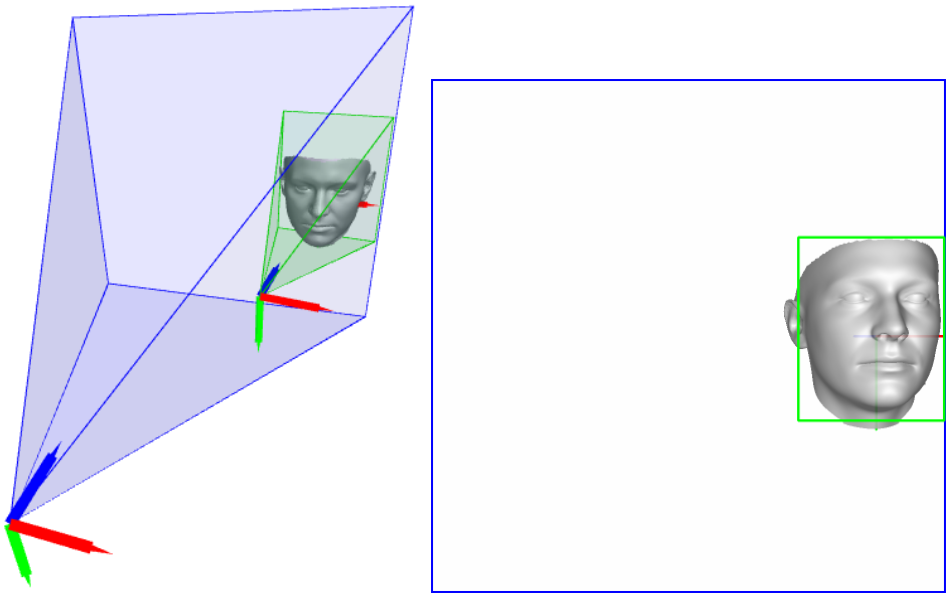


Figure 4.4: Illustration of the pose conversion of Algorithm 1. Left: Frusta of whole image camera C_{image} (blue), BBox camera C_{bbox} (green) and a head pose (mesh). Right: projection of the 3D scene into the camera image using K_{image} and a bounding box of the head (green). C_{bbox} is a virtual camera with a focal length proportional to its bounding box size, thus representing a canonical size. C_{bbox} is closer to the head pose and the principal axis goes through the bounding box center. As a result, a nearly frontal pose estimated within the bounding box will be converted to a head pose close to the right border of the whole image, rotated and further away from C_{image}.

This chapter defines the translation loss $L_{\text{t}}(t, \hat{t}) = ||t - \hat{t}||_2^2$ for the estimated translation t and the ground truth translation $\hat{t}$. The rotation loss $L_{\text{R}}(R, \hat{R}) = L_{\text{geodesic}} = \arccos\left(\frac{\text{tr}(R\hat{R}^T)-1}{2}\right)$ corresponds to the geodesic distance between the predicted rotation R and the ground truth $\hat{R}$. Optimization target is the overall loss L:

$$L = L_{\text{objectness}} + L_{\text{bbox}} + L_{\text{faceness}} + L_{\text{R}} + L_{\text{t}} \tag{4.3}$$

4.2.5. INTRINSICS-CONSISTENT IMAGE AND POSE AUGMENTATIONS

Augmentations are a scheme to create further training data to obtain a more robust model. The ground truth to the model is given by the tuple (image (h, w), camera intrinsics K, head pose $T^{\text{cam}\leftarrow\text{head}}$). The invariant of each augmentation is that the tuple remains consistent in the sense that the augmented head pose is being projected onto the corresponding locations of the augmented image using the augmented camera intrinsics. This chapter employs intrinsics-aware crop, scale, and flip augmentations.

Crop image with $bbox_{\text{crop}} = [u, v, w, h]$ needs a shift of the principal point for the

augmented camera intrinsics:

$$K_{\text{crop}} = \begin{bmatrix} f_x & 0 & c_x - u \\ 0 & f_y & c_y - v \\ 0 & 0 & 1 \end{bmatrix} \tag{4.4}$$

Head pose remains the same.

Scaling image height/width with factor s_h/s_w needs rescaling of focal lengths and principal point: $K_{\text{scale}} = \text{diag}(s_w, s_h, 1) \cdot K$.

Flip (left-right) flips the principal point c_x:

$$K_{\text{flip}} = \begin{bmatrix} f_x & 0 & w - c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4.5}$$

The head pose needs to be flipped on the yz-plane of the camera frame, which can be obtained by the following Hadamard product ($\circ$, piecewise multiplication) and keeps the transform right-handed:

$$T^{\text{cam}\leftarrow\text{head}}_{\text{flipped}} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \circ T^{\text{cam}\leftarrow\text{head}} \tag{4.6}$$

4.2.6. TRAINING DETAILS

The proposed intrApose model was implemented in PyTorch with a ResNet-18 backbone [50] which was pretrained on natural images. All implemented modules are differentiable to allow gradients to flow backward from the losses. This includes orthogonalization and pose conversion modules. Stochastic gradient descent (SGD) is used on mini-batches of four images with an initial learning rate of 0.001 and a weight decay of $5 \cdot 10^{-4}$. The learning rate was reduced by a factor of 10 if the model has not improved over the last three epochs on the validation set. Similarly, early stopping was performed after 5 epochs without improvement on the validation set.

For training the RPN, 256 bounding box proposals were sampled randomly per image. For training the faceness prediction net and the raw pose prediction net, 512 proposals per image were sampled.

The training data were augmented by intrinsics-aware scaling, mirroring, and cropping, as detailed in Section 4.2.5. Unbiasing: To make the model more robust in non-frontal poses, training samples with non-frontal poses are sampled more frequently compared to the dataset distribution which typically consists of more frontal driver head poses in in-car settings.

Training converged after 11 epochs and took approximately 2.5 days on a single NVidia Tesla V100 GPU. The model has $4.2 \cdot 10^7$ parameters. Inference time on the `float32` model is 18.4 samples per second.

4.3. EXPERIMENTS

Evaluation of intrApose puts specific requirements on the evaluation dataset. Datasets that do not provide camera intrinsics become out of scope as argued in Section 4.2.3. The evaluation is therefore based on *DD-Pose*, the large-scale in-car dataset that was introduced in Chapter 3 and depicts complex naturalistic driving scenarios. *DD-Pose* offers dataset splits depending on occlusion annotation and angle-from-frontal (all, easy, moderate, hard), see Section 3.2.7. Subjects {8, 19, 23} were chosen for validation, subjects {3, 6, 10, 11, 14, 15, 16, 17} for testing and the other subjects for training.

4.3.1. MODEL VARIANTS

Experiments were conducted with five different approaches, distinguished by model used (img2pose vs. intrApose), training dataset (WIDER vs. *DD-Pose*), and rotational representation (rotvec vs. SVDO$^+$). Recall, translation error, and rotation error are evaluated.

img2pose(WIDER, rotvec): Pretrained img2pose model provided by Albiero *et al.* [4] (see Section 2.1.2). The authors trained the model on the WIDER face dataset, which does not provide camera intrinsics. 3D head poses on the WIDER dataset were created by the authors using Perspective-n-Point on facial landmarks with a large head model and an assumed focal length equaling the sum of image width and image height, i.e., the same focal length assumption the method makes internally. One important fact to mention is that using the same focal length for the creation of the ground truth pose imposes a bias. The large head model (width $\approx 1.5\,\text{m}$, 10 times as large as a mean head) introduces head translations about 10 times as far away from the camera.

img2pose(DD-Pose, rotvec): This is the same model as *img2pose(WIDER, rotvec)*, but trained on *DD-Pose* using the training scheme of Albiero *et al.* [4], i.e., using assumed focal lengths instead of the true camera intrinsics provided with the *DD-Pose* dataset. Bounding boxes and head poses were used as provided by *DD-Pose*. 3D head landmarks of typical size (width $\approx 0.15\,\text{m}$) were needed in the calibration point loss of img2pose for the training to converge, potentially caused by points of the large model being projected outside the image in the calibration point loss. Note that the points used by calibration point loss are not to obtain a scale (as with landmark-based approaches), but rather to guide the model in adapting its parameters for pose estimation during training.

intrApose(DD-Pose, rotvec): The proposed model with the discontinuous rotation vector (rotvec) representation and the L_2 loss function of img2pose. This model is intrinsics-aware. Pose normalization was used as in img2pose, i.e., estimating the pose with zero-centered mean and unit standard deviation. A head model of typical size was used for the calibration point loss and applied the proposed intrinsics-aware crop, flip, and scale augmentations defined in Section 4.2.5.

intrApose(DD-Pose, SVDO$^+$): The proposed model with the continuous pose representation SVDO$^+$ [69] and geodesic loss. The model is intrinsics-aware. Pose normalization was found to be unnecessary. Anchor sizes and aspect ratios were tuned on the *DD-Pose* training set. Compared to img2pose, no point calibration loss was necessary. A head model of typical scale was used to create bounding boxes from the predicted head poses.

intrApose(DD-Pose, SVDO$^+$, unbiased): Same model as *intrApose(DD-Pose, SVDO$^+$)*, but trained with an unbiased dataset by sampling more non-frontal poses. The rationale is to make the model more robust in non-frontal poses.

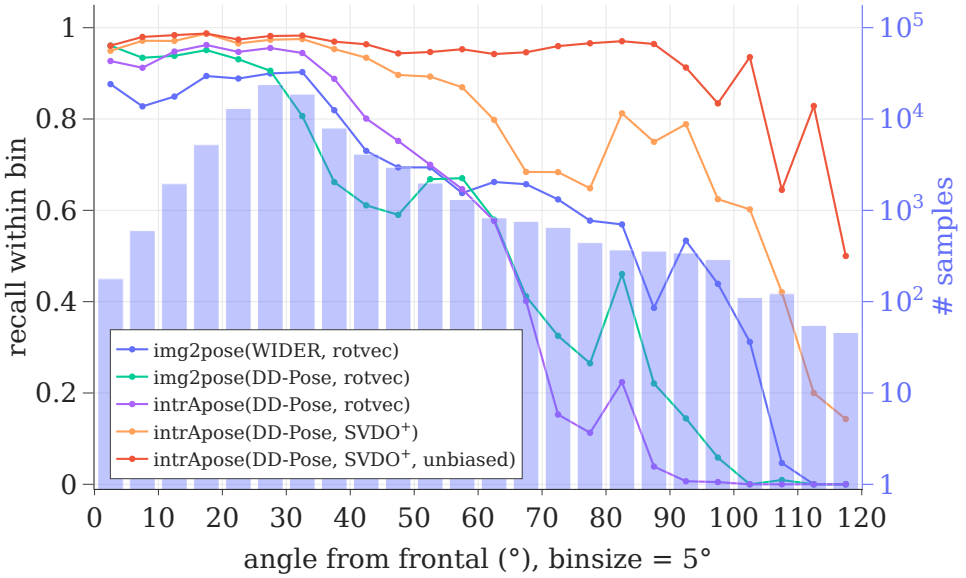


Figure 4.5: Recall and data distribution over the angular difference from frontal pose for the *DD-Pose* test set. Compared to the baseline (*img2pose(WIDER, rotvec)*), the recall improves incrementally by training on *DD-Pose* (*img2pose(DD-Pose, rotvec)*), switching to the proposed architecture (*intrApose(DD-Pose, rotvec)*), using a continuous rotation representation (*intrApose(DD-Pose, SVDO$^+$)*) and training using an unbiased dataset with more non-frontal poses (*intrApose(DD-Pose, SVDO$^+$, unbiased)*).

4.3.2. RECALL

Recall defines on which percentage of the images a head hypothesis from head pose estimation method exists. Images without a hypothesis are left out when evaluating translation and rotation. For matching ground truth and hypotheses, an Intersection over Union (IoU) threshold of 0.3 was used for the respective bounding boxes. Predicted head poses with a faceness score of > 0.9 are considered.

Figure 4.5 depicts the recall over the angle difference from frontal head pose. The rotation vector-based methods (rotvec) have a recall of > 0.8 for frontal faces and drop towards 0.6 for rotations 60° off-frontal. Out of these, the model variants trained on *DD-Pose* have a higher recall for close-to-frontal poses. The baseline *img2pose(WIDER, rotvec)* offers a higher recall for highly off-frontal faces ($[70°, 100°]$) compared to the other rotation vector-based methods. This can be explained by the WIDER dataset having a more homogeneous angular distribution compared to *DD-Pose*, which offers more close-to-frontal faces (see histogram in Figure 4.5).

Using the continuous SVDO$^+$ rotation representation (*intrApose(DD-Pose, SVDO$^+$)*) shows a considerable benefit across the whole angular spectrum compared to the rotation vector representation (*intrApose(DD-Pose, rotvec)*), keeping the recall above 0.6 for angles up to 105° and dropping towards 0.4 for 110°. The same model trained with an unbiased dataset (*intrApose(DD-Pose, SVDO$^+$, unbiased)*) shows remarkable improvement of recall for extreme poses, keeping the recall above 0.8 across the whole angular spectrum until 105°, only afterward dropping towards 0.4. The right side of Table 4.1 shows recall aggregated

Table 4.1: Rotation errors, translation errors and recall on the *DD-Pose* test set for the model variants on different subsets (all, e: easy, m: moderate, h: hard). See Section 4.3.1 for details on the models. Rotation errors are given in degrees (°), and translation errors in millimeters (mm). ↑/↓: higher/lower values denote better performance.

Method	BMAE (°) ↓				MAE_R (°) ↓				MAE_t (mm) ↓				recall (%) ↑			
	all	e	m	h	all	e	m	h	all	e	m	h	all	e	m	h
img2pose(WIDER, rotvec)	10.3	6.4	11.1	20.3	7.8	6.7	9.4	18.4	7849	7746	8068	8431	85	99	64	56
img2pose(DD-Pose, rotvec)	14.8	5.9	12.5	48.3	6.9	5.1	8.7	42.6	**18**	**14**	23	78	81	92	68	33
intrApose(DD-Pose, rotvec)	7.5	6.4	7.5	12.6	6.3	6.0	6.8	9.7	19	18	**21**	**26**	89	99	80	24
intrApose(DD-Pose, SVDO$^+$)	8.0	**4.0**	8.0	15.3	5.0	4.0	**5.9**	16.0	21	18	24	47	95	**100**	90	71
intrApose(DD-Pose, SVDO$^+$, unbiased)	**5.8**	4.2	**6.2**	**9.5**	**4.8**	**3.9**	**5.9**	**8.9**	25	22	29	41	**97**	**100**	**93**	**93**

over the subsets (all, easy, moderate, hard) in accordance with the observations from Figure 4.5.

4.3.3. TRANSLATION ERROR

For translation error, the mean Euclidean distance mean absolute error (translation) (MAE_t) between ground truth head origin and predicted head origin is used.

The errors in head translation estimation (MAE_t) are listed in Table 4.1. The pre-trained baseline *img2pose(WIDER, rotvec)* depicts an error of over 7.7 m. Overestimated distance to camera (t_z) contributes most to the error. This is caused by two facts: for one, *img2pose(WIDER, rotvec)* assumes a focal length defined by the image size which does not correspond to the true intrinsics of the camera. Also, the WIDER dataset consists of 2D facial landmark labels which Albiero *et al.* use to generate the ground truth head poses by Perspective-n-Point and a 3D head model which is ~ 1.5 m wide. The model trained on WIDER therefore estimates heads presented in *DD-Pose* further away. As Albiero *et al.* use WIDER for both training and evaluation, this fact had not become apparent. Albiero *et al.* emply the same large head model both for obtaining ground truth (Perspective-n-Point) and within their model. This leads to a biased comparison in [4], as the evaluation revealed by using a dataset that provides intrinsics and does not create ground truth using the same assumptions of intrinsics.

In comparison, the img2pose-based model trained on *DD-Pose* (*img2pose(DD-Pose, rotvec)*) shows a better estimation of the head translation, caused by the correct head pose ground truth obtained by a measurement device. Overall, the head is estimated 18 mm from the ground truth for the *all* subset and 78 mm for the *hard* subset. The non-unbiased intrApose model variants perform similarly in translation estimation, being less than 21 mm off for the *all* subset. Comparing the SVDO$^+$ model variants shows that unbiasing the training dataset decreases translation error from 47 mm to 41 mm on the *hard* subset, though sacrificing MAE_t for the other subsets (*easy*, *moderate*). The worsening on the latter subsets can be explained by the use of significantly fewer training samples from these subsets while evaluation is biased in the sense that a large portion of samples resides in the *easy* and *moderate* subsets (see data distribution in Figure 4.5). Another hypothesis is a higher imbalance of rotation loss and translation loss (unbiasing is based on angle from frontal).

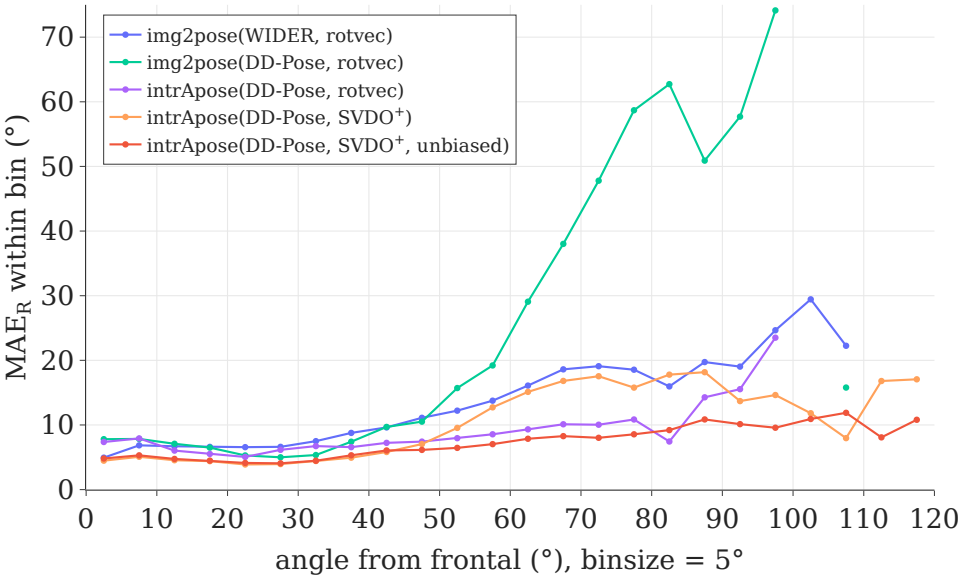


Figure 4.6: Mean absolute error (rotation) (MAE_R) on the *DD-Pose* test set. MAE_R increases with larger angular distance from frontal. The model variant *intrApose(DD-Pose, SVDO$^+$, unbiased)* provides a consistently low angular error over the whole depicted angular spectrum.

4.3.4. ROTATION ERROR

Rotation error is evaluated by mean absolute error (rotation) (MAE_R) of the geodesic distance between ground truth rotation and predicted rotation. For an unbiased evaluation of head rotation, this chapter uses balanced mean absolute error (BMAE) as proposed by Schwarz *et al.* [119]. It splits the dataset in bins based on the geodesic distance from the frontal pose and averages the MAE_R of the bins:

$$\mathrm{BMAE}_{d,k} := \frac{d}{k} \sum_i \phi_{i,i+d} \; \forall i \in d\mathbb{N} \cap [0,k] \tag{4.7}$$

where $\phi_{i,i+d}$ is the MAE_R of all hypotheses, where the geodesic distance between ground truth and frontal pose is between i and $i+d$. During evaluation, bin size $d := 5°$ and maximum angle $k := 120°$ are used.

The overall MAE_R and BMAE are displayed in Table 4.1 and the MAE_R over the angular difference from a frontal pose are depicted in Figure 4.6.

The pretrained baseline *img2pose(WIDER, rotvec)* shows a MAE_R/BMAE of 7.8°/10.3° on the *all* subset of *DD-Pose*, though being trained on WIDER, a dataset based on images downloaded from the internet, therefore shows good generalization to unseen data.

Retraining the img2pose model on *DD-Pose* (*img2pose(DD-Pose, rotvec)*) decreased the MAE_R to 6.9°, yet increasing the BMAE to 14.8°. This is due to the worse performance for non-frontal poses (see increasing MAE_R with increasing angle-from-frontal in Figure 4.6). This can be explained by the majority of the training samples within *DD-Pose* being close-to-frontal, making the model tend to estimate the mean pose with the discontinuous rotation vector representation.

intrApose(DD-Pose, rotvec) uses the same data and pose representation within the proposed intrinsics-aware intrApose framework including the proposed augmentations. The BMAE decreases to 7.5° on the *all* subset and considerably improves on the *hard* subset to 12.6° (from 20.3° and 48.3° of the img2pose models trained on WIDER and *DD-Pose*, respectively). This improvement can be attributed to the intrinsic-aware model.

Switching to the continuous rotation representation SVDO$^+$ (*intrApose(DD-Pose, SVDO$^+$)*) decreases the MAE_R to 5.0°, yet increases in terms of BMAE (8.0°). A look at the corresponding recall on the *hard* subset shows that it now predicts more extreme poses (71% vs. 24%) and still tunes towards close-to-frontal poses, as shown by the best MAE_R on the *easy* subset. Overall, one can say that the closer the BMAE of a model is to the MAE_R, the better it covers data-imbalance.

The final, proposed model *intrApose(DD-Pose, SVDO$^+$, unbiased)* resolves this data imbalance by being trained with more off-frontal pose samples. An improvement of both MAE_R and BMAE can be seen on the hard subset of *DD-Pose*. The model variant shows a consistently low error along the full spectrum of angles from frontal (Figure 4.6) and results in a BMAE of 9.5° on the *hard* subset, being very close to the corresponding MAE_R of 8.9°.

Experiments with the continuous Ortho6D rotation representation of Zhou *et al.* [144] showed results similar to the SVDO$^+$ rotation representation, in accordance with the observations of Levinson *et al.* [69].

4.3.5. QUALITATIVE RESULTS

Figure 4.7 provides qualitative results of the baseline img2pose [4] and the proposed model. The small axes of img2pose confirm the overestimated head translation observed in Table 4.1. As designed, the unbiased proposed model depicts a smaller qualitative error for off-frontal poses compared to the unbiased variant. Samples where only the proposed model could provide a head pose estimate (faceness > 0.9) are depicted in Figure 4.8. The model shows robustness towards high occlusions by hands and steering wheel, and extreme poses, though with a larger qualitative error compared to the samples given in Figure 4.7.

4.4. DISCUSSION

This chapter presented a 6 DOF head pose estimation method which employs a continuous rotation representation. For more than two decades, authors have committed to Euler angles or quaternions and treated the values as a simple regression problem, thus ignoring the underlying manifold at hand. This led to complex mitigations dealing with the drawbacks of the representations, such as coarse-to-fine approaches, normalizing values (zero mean, unit standard deviation, quaternion normalization), explicitly handling discontinuities (e.g., at 360°) or proposing special losses (e.g., by encouraging orthogonality or projection of calibration points). This chapter confirmed the importance of the proper choice of rotation representation of Levinson *et al.*[69]: 3 DOF rotation could be represented without special pre- or postprocessing, thus leading to a plain network without bells and whistles.

When evaluating, representing rotation errors by Euler angles shows drawbacks due

to their ambiguity (order of axes, direction, handedness). Therefore, this chapter employs geodesic distance, making it agnostic of single angle components and the frame, but at the cost of lacking insight into the contribution of individual axes to the geodesic distance.

The proposed architecture is based on the Faster R-CNN framework. In general, it can be adapted to other, potentially deeper backbones or to single-shot detection networks.

The method is intrinsics-aware, therefore requiring camera parameters alongside the image itself. However, Section 4.2.3 also showed that assuming incorrect camera intrinsics can introduce large errors beyond accepted tolerances. With missing calibration information the error behavior of the proposed method assimilates to methods that are intrinsics-agnostic.

The pose prediction network operates on feature RoIs, therefore can only estimate a local pose within the BBox. Albiero *et al.* [4] proposed a pose conversion to the whole image. This chapter generalized the pose conversion algorithm by making it intrinsics-aware, allowing for generic pose representations, and explicitly formalizing a necessary orthogonalization step. Essentially, this shows that the pose conversion is a rigid coordinate transformation that approximates the projections of the BBox and the whole image.

Rendering a face mesh overlay is an appealing visualization of head pose. However, using a fully opaque one makes the viewer tolerate more errors both in rotation and translation, by still appearing natural. To that end, rendering the face mesh transparent and also visualizing the frame axes is suggested.

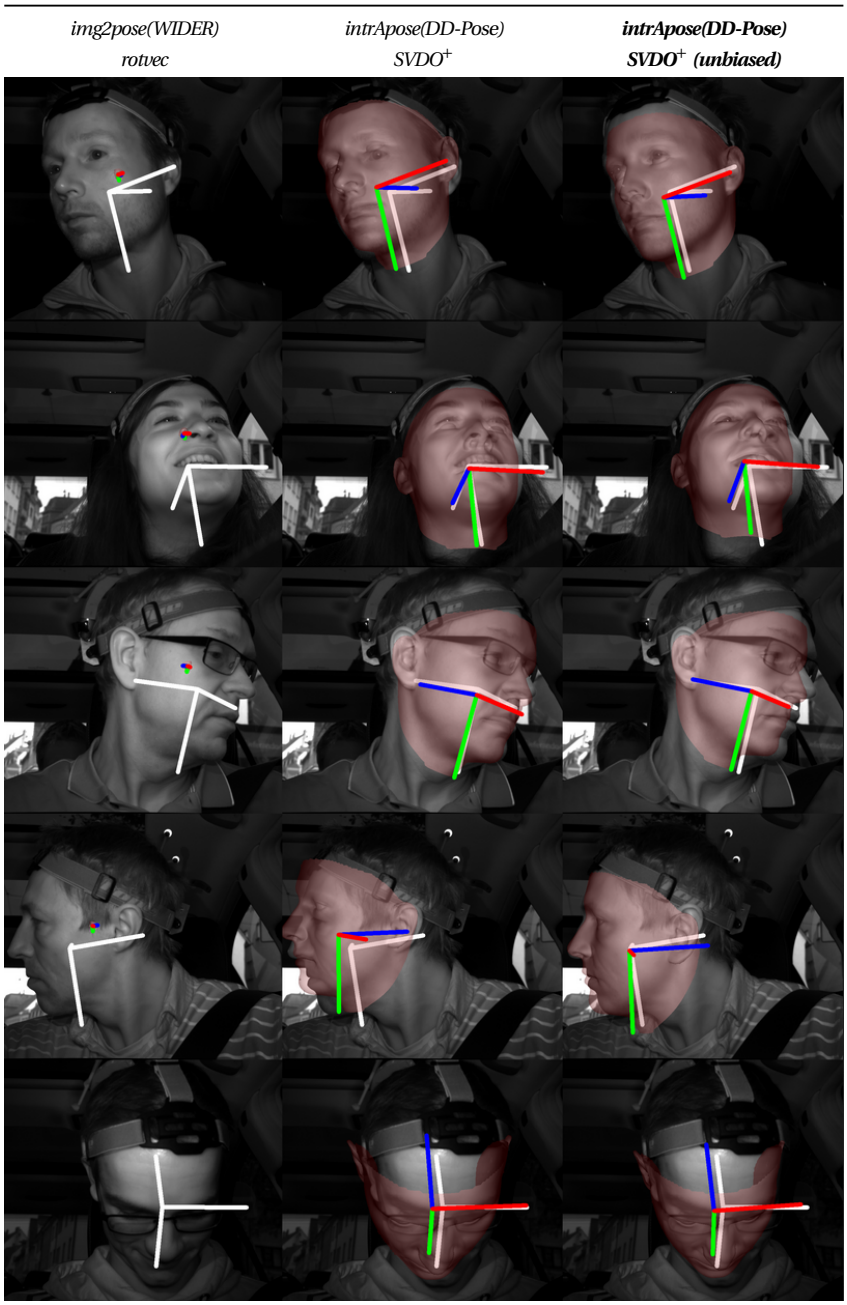


Figure 4.7: Qualitative head pose estimation results on samples with challenging off-frontal head poses. The poses are projected into the camera image using the camera intrinsics. Ground truth head pose is denoted by a white axis. Predicted head pose is denoted with an RGB axis and a transparent red face mesh of typical head size. The translation error can be judged by comparing the axis length. Note that the small axes of *img2pose(WIDER, rotvec)* are caused by overestimation of the distance-to-camera. All images are crops to ease judgment.

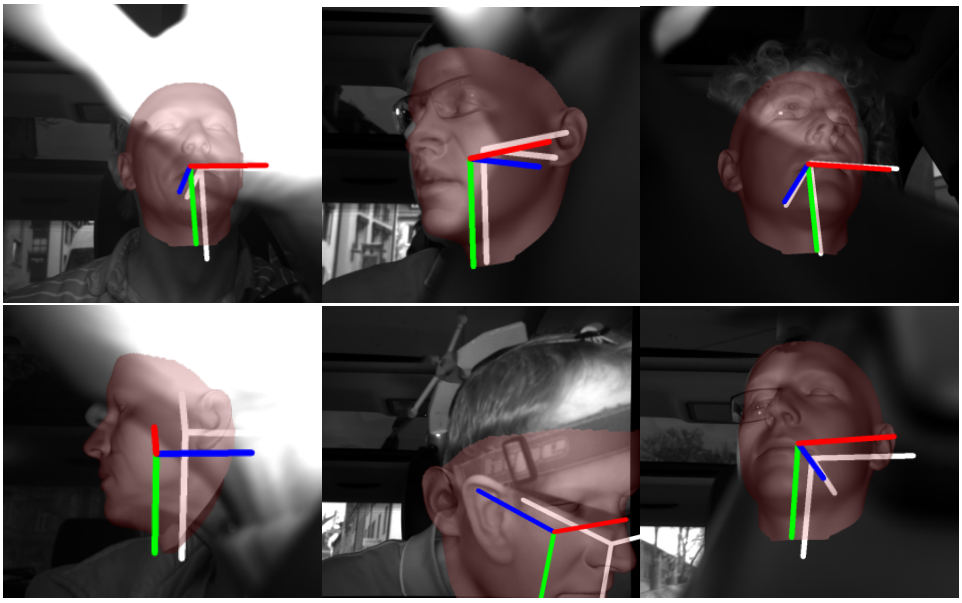


Figure 4.8: Random subset of samples of the *DD-Pose* test set where a head pose estimate could only be provided by the proposed model (unbiased). They show high occlusions by hands and steering wheel, and extreme poses. All images are crops to ease judgment.

5

Deep End-to-end 3D Person Detection from Camera and Lidar

This chapter presents a method for localizing persons in the surrounding of an intelligent vehicle. The 3D location of a person is a key information for modeling the interaction with the driver/ego-vehicle in Chapter 6.

5.1. Objectives

In intelligent transportation systems, different sensors are commonly used for person detection, such as cameras, radar and lidar sensors, each coming with their individual advantages and drawbacks. While a monocular camera-based system offers a dense projection of the light in the field of view, it lacks distance information. Lidar sensors offer a sparse scan of the environment with precise distance information, even during nighttime. However, even modern lidar sensors offer only 128 vertical layers.

In the last two decades, person detection performance has significantly improved due to machine learning methods and the rise of large and representative datasets to optimize and evaluate the methods on [12, 43]. In the past years, deep learning methods have shown to be very accurate in the task of *2D* person detection in camera images, which in contrast to *3D* person detection, does not estimate the distance of a person to the camera sensor. 3D person detection remains more challenging, compared by the detection performance on standard benchmarks, such as the KITTI 3D Object Detection Evaluation [43]. Methods performing solely on 3D points clouds, or additionally taking camera images as an input still perform mediocre with an average precision (AP) of 46.6% on the moderate subset of the KITTI benchmark. For multi-modal methods, the problem of sensor fusion has to be solved. Some methods rely on transforming the point cloud to an image-alike structure, such as bird eye view (BEV), to employ methods known from image recognition.

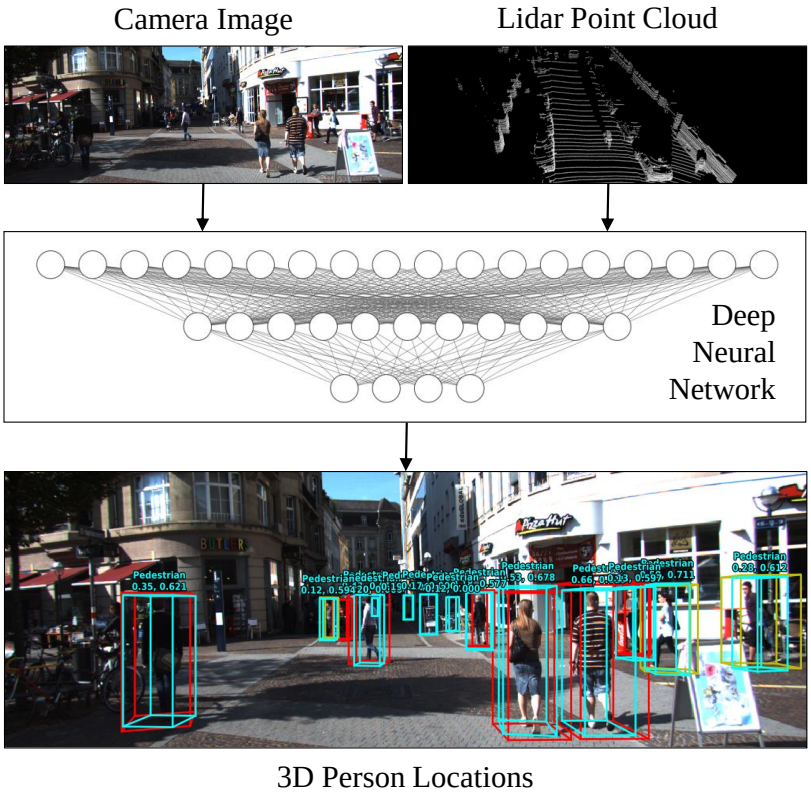


Figure 5.1: The proposed method estimates 3D person bounding boxes from camera images and 3D lidar point clouds. The deep end-to-end model learns low-level features for both modalities, fuses their higher-level representations, and predicts the 3D location of the persons in the scenes. Predictions are depicted by cyan bounding boxes which are projected in the camera image. Ground truth bounding boxes are shown in red. The numbers on top of each prediction denote (objectness, Intersection over Union (IoU)). Evaluations are performed on the KITTI dataset [43].

Hand-crafted preprocessing of lidar point clouds raises questions such as “are there representations which perform better?". This question is less likely to be raised for end-to-end learning, where sensor input is kept as raw as possible to have representations being *learned* in contrast to being *designed*.

This chapter presents a deep learning based method for 3D person detection, which performs end-to-end learning on sensor data from camera and lidar. High-level representations from image and point cloud are learned from the raw sensor data. For the image input, a VGG-like convolutional neural network extracts a high-level feature representation. For the point cloud input, a Voxel Feature Encoder (VFE) is employed for abstract feature extraction. Features from both modalities are fused to serve as an input for a regression model which estimates the 3D positions of persons. See Figures 5.1 and 5.2 for an overview of the proposed system.

This chapter is based on the work published in [110] (©2019 IEEE).

5.2. PROPOSED APPROACH

This chapter presents an end-to-end method for 3D person detection based on camera images and lidar point clouds [56]. The proposed approach builds upon the architecture of Aggregate View Object Detection (AVOD) [63]. As in AVOD, feature maps from both a camera and lidar modality are extracted. A region proposal network (RPN) generates 3D region proposals based on cut-outs of the feature maps of 3D anchors. The top region proposals are refined by a second stage detection network which estimates the 3D location and spatial extent (i.e., length, width and height) of the persons present in the scene.

In contrast to AVOD, which relies on hand-crafted bird-eye-view (BEV) for the lidar input, the presented approach learns point cloud features by applying Voxel Feature Encoding (VFE) layers followed by 3D convolutional layers for high level feature extraction as introduced in VoxelNet [146]. The proposed architecture is depicted in Figure 5.2.

5.2.1. INPUT PREPROCESSING AND FEATURE EXTRACTION

Both camera images and lidar point clouds are preprocessed to allow for subsequent feature extraction.

IMAGE PREPROCESSING

The RGB camera images are normalized by subtracting the mean RGB value of the training dataset. For image feature extraction, the VGG16 architecture [125] with the same modifications as in [63] is used, i.e. half the number of filters in each convolutional layer, no fifth convolutional stage and no max-pooling layer at the end of the fourth stage. The resulting 256 feature maps are eight times smaller along each dimension. To attain higher resolution feature maps, four times bilinear upsampling is applied.

POINT CLOUD PREPROCESSING

The lidar point cloud is cropped to reside in the volume $\Delta X = [-40\,\text{m}, 40\,\text{m}]$, $\Delta Y = [-1\,\text{m}, 3\,\text{m}]$, $\Delta Z = [0\,\text{m}, 70\,\text{m}]$ in the camera frame. The volume is partitioned into equally sized voxels of size $(s_x, s_y, s_z) = (0.2\,\text{m}, 0.4\,\text{m}, 0.2\,\text{m})$. The voxelized input is processed by a Feature Learning Network, as proposed in [146]. It consists of grouping, random sampling and stacked voxel feature encoding (VFE) layers. Each point $p_i = (x_i, y_i, z_i)$ in each voxel v forms an input vector $(x_i, y_i, z_i, r_i, \tilde{x}_v, \tilde{z}_v, \tilde{z}_v)$ with point reflectance r_i and voxel mean coordinates $(\tilde{x}_v, \tilde{z}_v, \tilde{z}_v)$. Each VFE layer learns a locally aggregated feature by a fully connected layer on the input, followed by max-pooling and concatenation [146]. $T = 45$ points per voxel are randomly sampled, followed by a stack of two VFE layers. The first VFE layer yields a 32 dimensional feature vector per voxel. The second VFE layers yields a 128 dimensional feature vector per voxel.

Three 3D convolutional layers conv3D(c, k, s, p) with output channels c, kernel size k, stride $s = (s_x, s_y, s_z)$ and padding $p = (p_x, p_y, p_z)$ aggregate voxel-wise features to obtain an expanding receptive field to capture more context information [146]. The parameters of the convolutional layers are:

$$\text{conv3D}_1\,(64, (1,2,1), (1,1,1))$$
$$\text{conv3D}_2\,(64, (1,1,1), (1,0,1))$$
$$\text{conv3D}_3\,(64, (1,2,1), (1,1,1))$$

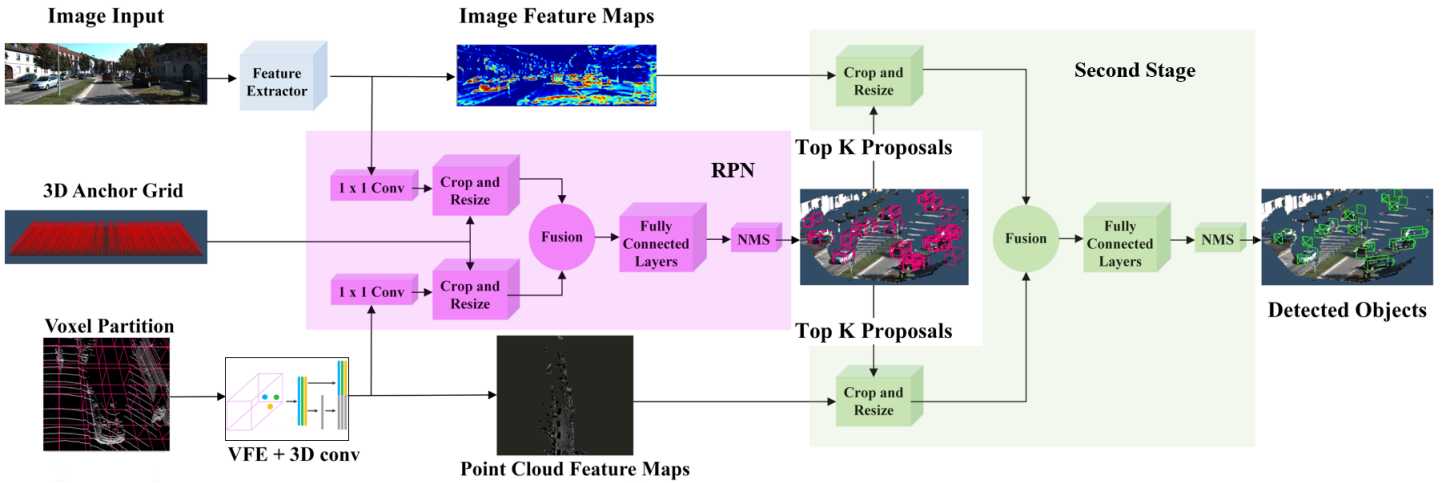


Figure 5.2: The architecture of the proposed method [56]. It is inspired by AVOD [63]. Image features are extracted by an adapted VGG16 network. Point cloud features are extracted from voxel partitions by applying Voxel Feature Encoding (VFE) layers and 3D convolutions. In a Region Proposal Network (RPN), 1×1 convolution is applied to the feature maps to reduce their size. Anchors from a 3D anchor are projected into the feature maps to crop proposals. After resizing to a common size, the feature crops from both modalities are fused and a the location of the objects is refined by a fully connected neural network. In the second stage, the best proposals from the RPN are cropped from the full feature maps and fused. Object detection layers are implemented by fully connected layers operating on the fused crops. This allows for an end-to-end network which estimates the 3D locations of persons from camera and lidar sensor data. Image adapted from [63, 146].

Reshaping is performed such that neighboring voxels along y dimension are flattened, thus resulting in a voxel feature map with 128 channels per voxel.

ANCHOR GENERATION

As in [63], anchors are spawned in a 3D dense grid in the voxel volume, using an interval of 0.5m along x and z direction and y coordinate to reside on the ground plane. The spatial extents of the anchors are based on size clusters obtained for each class on the training set. Anchors which are outside the camera view or not supported by any point are discarded.

5.2.2. REGION PROPOSAL NETWORK

The region proposal network (RPN) projects anchors to the feature maps of each modality, crops the respective feature residing in the anchor projections, resizes them and fuses them. Subsequently, fully connected layers refine the location of anchor boxes towards ground truth location to form the region proposals. The RPN is adopted from [63], but in contrast the presented approach crops from the learned voxel feature map instead of the bird eye view feature map. For the RPN, the feature maps are reduced in dimensionality by performing a 1×1 convolution [63] which can be seen as a learned weighting of all feature maps along the y dimension.

Crops of size 3×3 are extracted from the feature maps of each modality and the 1024 best proposals are obtained after non-maximum suppression. The crops are fused using the mean operation. There are two fully connected layers with 2048 neurons each.

Proposals to optimize are selected by having an Intersection over Union (IoU) of > 0.8 with the projected ground truth box. The proposed method uses a Smooth L1 loss for localization regression task and a cross-entropy loss for the classification task (person vs. background).

5.2.3. DETECTION NETWORK

The second stage detection network is also based on [63], i.e. region proposals are cropped from the feature maps of both modalities, fused and used to regress location, spatial extent and class by fully connected layers.

Crops of size 7×7 of the feature maps of both modalities are concatenated. There are three fully connected layers with 2048 neurons each.

The location and spatial extent of the detected persons are retained after a non-maximum suppression.

5.2.4. FUSION SCHEMES

Both the RPN and the detection network fuse resized feature map crops from both modalities. MV3D [23] proposes three different fusion schemes, namely *early*, *late*, and *deep* fusion. Combinations of individual input can be *concatenation* or *element-wise mean*.

The fusion schemes differ in which order feature transformations (e.g. convolutions) are applied compared to feature combinations. *Early fusion*: combine individual inputs, then transform. *Late fusion*: transform inputs, then combine. *Deep fusion*: combine inputs, then transform individually, and repeat. In deep fusion, the transformations of each repetition learn different parameters.

Table 5.1: Statistics of the KITTI dataset used for evaluation. Each scene represents a synchronized snapshot of the environment at a point in time. The number of 3D annotations in the camera's field of view is listed.

Dataset	# Scenes	# 3D Annotations		
		Cars	Cyclists	Pedestrians
KITTI train	3712	14357	734	2207
KITTI val	3769	14385	893	2280
KITTI test	7518	-	-	-

5.2.5. Training

The proposed end-to-end network is optimized using ADAM [59]. One scene per training iteration is used, yielding 1024 proposals for training the network. The RPN and detection network are trained jointly starting from a random initialization. The learning rate is set to 0.0001.

5.3. Experiments

The proposed method is evaluated on the pedestrian class of the KITTI 3D Object Detection Evaluation 2017 (KITTI) [43] using average precision (AP), which follows the standard evaluation protocol of the KITTI benchmark.

5.3.1. Dataset

The KITTI dataset [43] captures 15k urban traffic scenes by camera images and lidar-based point clouds. Traffic participants such as cars, cyclists and pedestrians are annotated by 3D bounding boxes. Three difficulty levels are defined (easy, moderate, hard) based on object size in the camera image, occlusion state and truncation ratio.

The annotated scenes are divided into a training split (*train*) and validation split (*val*), as in [22], which ensures that images from the splits are from disjoint sequences. See Table 5.1 for an overview of the KITTI dataset and the provided annotations.

5.3.2. Data Augmentation

As the KITTI *train* split offers around 4 k scenes with around 2.2 k pedestrians, augmentations are used to increase diversity in the training set. The image is flipped along the vertical axis and the corresponding point cloud along the yz plane. The principal point of the camera matrix is adapted accordingly to ensure valid projections into the flipped camera image.

5.3.3. Evaluation Metrics

The performance of the 3D object detection task is evaluated using average precision (AP), as defined in [38, 43], i.e.

$$\mathrm{AP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1.0\}} \max_{\mathrm{recall}(c) \geq r} \mathrm{precision}(c) \tag{5.1}$$

Table 5.2: Selected experiments of hyper parameters and their performance on the *val* split of KITTI.[1]

Exp.	RPN Stage		Detection Network			Ped. AP (%)		
	Comb.	Crop	Fus.	Comb.	Crop	Easy	Mod.	Hard
#1	mean	3×7	early	mean	7×7	45.85	40.79	35.92
#2	mean	3×3	deep	mean	7×7	44.18	37.11	30.36
#3	mean	3×3	late	mean	7×7	49.56	43.68	38.36
#4	mean	5×5	early	mean	9×9	50.00	44.47	38.70
#5	concat	3×3	early	mean	7×7	51.91	46.38	40.86
#6	mean	3×3	early	concat	7×7	53.29	46.23	40.28
#7	mean	3×3	deep	concat	7×7	**53.47**	**47.06**	**41.49**

with $\text{recall}(c) = \frac{\text{tp}(c)}{\text{tp}(c)+\text{fn}(c)}$, and $\text{precision}(c) = \frac{\text{tp}(c)}{\text{tp}(c)+\text{fp}(c)}$, both for an objectness confidence threshold c. $\text{tp}(c)$ and $\text{fn}(c)$ denote the number of true positives and false negatives, respectively.

A 3D prediction bounding box is considered to correspond to a 3D ground truth annotation bounding box, if the Intersection over Union (IoU) in xz coordinates is above 0.7. This follows the evaluation protocol of the KITTI 3D Object Detection Evaluation 2017 [43].

5.3.4. EXPERIMENTAL RESULTS

The *train* split is used to train the proposed model using different hyper parameters and evaluate on the *val* split to evaluate the performance using AP. The hyper parameters under test were fusion schemes and combination methods for both the RPN and the second stage detection network, as introduced in Section 5.2.4. Additionally, the feature crop size was varied.

QUANTITATIVE ANALYSIS

All models were trained starting from random initialization and continuously evaluated on the *val* split every 1000 training iterations, stopping at maximum 120k iterations. For all comparisons, the best-performing model over all training iterations was chosen.

Table 5.2 shows selected experiments whichwere conducted and their respective performance on the *val* split. Using the late fusion scheme in the detection network yields a higher performance than deep or early fusion when combining the individual features via an element-wise mean (experiments #1 - #3). Increasing the crop sizes increases AP when keeping element-wise mean (experiment #1 vs. #4). The highest performing experiments are using concatenation feature combination in the detection network (experiments #6 & #7). Among those experiments the deep fusion scheme

[1] As typical for high dimensional models, alterations in the hyper parameter space have a highly non-linear impact on the model performance. Therefore only a subset of the conducted experiments with good performance is shown.

Table 5.3: Comparison of the proposed method vs. SoA at the time of publication [110]. Results on the KITTI 3D detection benchmark. Test split is used, unless otherwise given. Numbers are given in % average precision (AP) for IoU > 0.7.

Method	Modality	Pedestrian		
		Easy	Mod.	Hard
F-PointNet [95]	lidar	51.2	44.9	40.2
PointPillars [65]	lidar	52.1	43.5	**41.5**
VoxelNet [146]	lidar	39.5	33.7	31.5
AVOD [63]	lidar & image	38.3	31.5	27.0
Proposed (*val*)	lidar & image	**53.5**	**47.1**	**41.5**

performs slightly better than early fusion scheme. Therefore the model from experiment #7 was chosen for further experiments.

Table 5.3 shows quantitative performance of the proposed model compared to selected State-of-Art (SoA) methods on the KITTI 3D Detection Benchmark. F-PointNet [95] and PointPillars [65] were chosen as they are among the best performing methods on the KITTI benchmark[2]. VoxelNet [146] and AVOD [63] are of special interest for comparison, as the former uses the same feature extraction for point cloud features and the latter introduced the architecture the proposed method is based on.

As can be seen, the proposed method outperforms the baselines on all three difficulties.

QUALITATIVE ANALYSIS

3D detection examples are presented in Figure 5.3. The detected 3D bounding boxes are projected into the camera images.

Figures 5.3a to 5.3d represent examples with good detection performance. The proposed model detects persons in crowded scenes (Figure 5.3a), as well as far away objects (Figure 5.3b). The model is robust against occlusions and truncations on the image border (Figures 5.3c and 5.3d).

Limitations of the system are presented in Figures 5.3e to 5.3h. While the system detects some highly occluded persons, it misses others and creates false positives (Figure 5.3e). Figure 5.3f shows a traffic structure which is falsely detected as a person. Figures 5.3g and 5.3h show confusions with cyclists.

[2] http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=3d, retrieved 2019-04-09

Figure 5.3: Qualitative results for the proposed model on the KITTI validation set. The model can estimate the 3D location of people with different orientations and poses. (a) - (d): examples with good performance on crowded scenes, far away objects, occluded and truncated objects. (e) - (h): failure cases with missed detections, false positives and confusions with cyclists. Ground truth: red. Predictions: cyan. Numbers on top of boxes: detector objectness, IoU.

6

Driver and Pedestrian Mutual Awareness for Path Prediction and Collision Risk Estimation

Chapter 4 and Chapter 5 introduced two building blocks by looking-in (driver head pose estimation) and looking-out (person detection) of the ego-vehicle. This chapter connects these building blocks towards the overall goal of this thesis and presents a novel method for path prediction of ego-vehicle and a pedestrian. It takes both head-pose of the driver and a pedestrian into account to estimate the mutual awareness towards each other.

6.1. Objectives

Maneuvering urban environments consists of an abundance of interactions between the driver and other traffic participants, such as pedestrians. Understanding these driver-pedestrian interactions is key for early and robust prediction of the future paths of the vehicle driven by the driver and pedestrians, allowing for assessing critical situations and initiating counter-measures, such as emergency brakes.

This chapter considers the setting of a potentially crossing pedestrian and an approaching vehicle that has the right-of-way (i.e. no dedicated crossing location). A method is presented which uses context cues about the spatial environment, driver-pedestrian mutual awareness, and potential motion coupling to estimate the future paths of both participants and associated collision risk. See Figure 6.1 for an illustration of the overall system.

Specifically, this chapter extends the Dynamic Bayesian Network (DBN) method from Kooij *et al.* [60, 61], which performs path prediction for an individual pedestrian, to the mutual vehicle-pedestrian case. As in [60, 61], it is captured that pedestrian awareness of the on-coming vehicle will likely affect his/her future path. The proposed method also models that driver awareness of the pedestrian will likely affect the future ego-vehicle

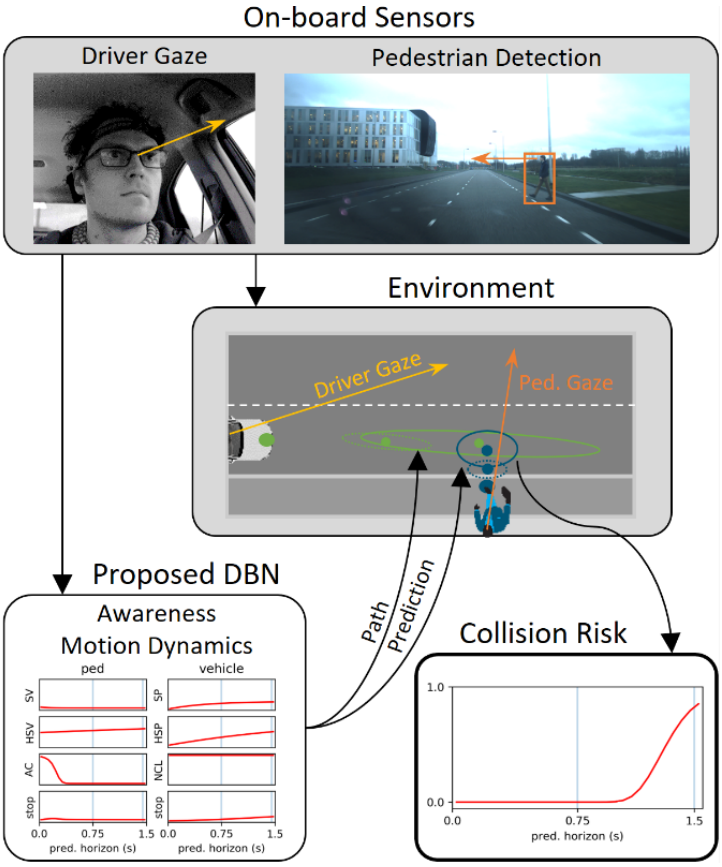


Figure 6.1: The system assesses mutual awareness of pedestrian and driver in a scenario of a potentially crossing pedestrian. Cues about the driver, pedestrian, and spatial environment are collected from on-board sensors. A probabilistic framework based on a Dynamic Bayesian Network (DBN) estimates latent states of awareness of the driver and pedestrian to predict their future motion. Based on the predicted paths, collision risk is estimated.

path. It uses head pose (pedestrian, driver) and eye gaze (driver) as proxies for awareness, as the latter cannot be determined directly.

There are several reasons for choosing a physics-based DBN approach for path prediction, as opposed to the popular neural networks. First, a DBN allows more easily to incorporate expert domain knowledge by means of its graphical model structure. Second, a DBN is interpretable, one can inspect the values of its latent variables and follow how it reaches its output. This is especially important for safety-critical applications. Third, one can expect a DBN to deal well with smaller datasets, as it has a comparatively small set of parameters, which will minimize the effects of over-training. Finally, recent work by Pool *et al.* [93] suggests that a DBN can deliver competitive path prediction results compared to a RNN, when its parameters are optimized by backpropagation as well.

This chapter is based on the work published in [111] (©2022 IEEE).

6.2. JOINT VEHICLE AND PEDESTRIAN PATH PREDICTION

Kooij *et al.* [61] note that a pedestrian's decision to continue walking or to stop in a crossing scenario is mainly influenced by the presence of an approaching vehicle on collision course, the pedestrian's awareness thereof, and the position of the pedestrian with respect to the curbside. This knowledge is encoded in a context-based Switching Linear Dynamical System (SLDS) (a special DBN), where latent discrete states control the switching probabilities between the continuous states dynamics of walking and standing.

This chapter covers vehicle-pedestrian collision risk, thus extends the prediction component to the ego-vehicle. Analogously one can argue that the vehicle's outcome of continue moving or stopping is mainly influenced by the presence of an approaching pedestrian on collision course, the driver's awareness thereof and the distance of the vehicle to the pedestrian's crossing location. Pedestrian and vehicle motion is modeled with two SLDSes which are linked to each other by a shared latent state, that captures the motion coupling between the two objects. The proposed DBN is shown in Figure 6.2 (see Table 6.1 for the corresponding node descriptions).

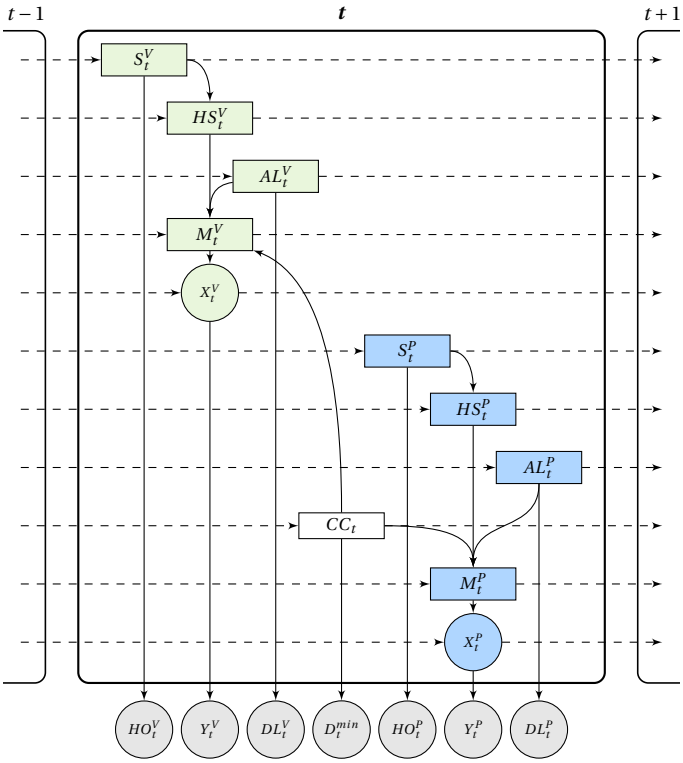


Figure 6.2: Graphical model representation of the Dynamic Bayesian Network (DBN). Discrete nodes are rectangular, continuous nodes are circular. Grey nodes represent observable variables while the other nodes represent latent states. Dashed lines depict temporal connections between latent context states in subsequent time instances. Driver-related nodes are shaded in green while pedestrian-related nodes are shaded in blue. Context state description and purpose are provided in Table 6.1.

Table 6.1: Latent context states, their associated observation and the purpose within the DBN structure. States are grouped by vehicle/driver (common superscript V), pedestrian (superscript P) and shared contexts.

Latent State	Abbr.	Observation	Abbr.	Purpose
driver-sees-pedestrian	S^V	driver-head-orientation (gaze)	HO^V	encodes driver's awareness of the pedestrian
driver-has-seen-pedestrian	HS^V	-	-	memorizes driver's (past) awareness of the pedestrian
vehicle-at-location	AL^V	vehicle-distance-to-location	DL^V	manifests typical location of braking (ped. crossing location)
vehicle-motion-model	M^V	-	-	switches between *driving* and *braking* LDS
vehicle-position-state	X^V	vehicle-position	Y^V	LDS for vehicle state estimation
pedestrian-sees-vehicle	S^P	pedestrian-head-orientation	HO^P	encodes pedestrian's awareness of the driver/vehicle
pedestrian-has-seen-vehicle	HS^P	-	-	memorizes pedestrian's (past) awareness of the driver/vehicle
pedestrian-at-location	AL^P	pedestrian-distance-to-location	DL^P	manifests typical location of stopping (curb)
pedestrian-motion-model	M^P	-	-	switches between *walking* and *standing* LDS
pedestrian-position-state	X^P	pedestrian-position	Y^P	LDS for pedestrian state estimation
collision-course	CC	minimum-future-distance	D^{min}	separates early crossings from critical crossing

6.2.1. DBN

The DBN consists of two sub-graphs, one for the pedestrian and one for the vehicle. The pedestrian sub-graph is congruent with the DBN of Kooij *et al.* [61]. The vehicle sub-graph displays analogous behavior for the vehicle, by encoding driver awareness by driver gaze and braking manifestation by being close to the crossing location of the pedestrian.

PEDESTRIAN-RELATED CONTEXT STATES

The pedestrian P can exhibit one of two motion types: *walking* ($M_t^P = m^P_{\text{move}}$, constant velocity) and *standing* ($M_t^P = m^P_{\text{stop}}$, constant position). The motion state of the pedestrian contains two-dimensional positions and velocities: $X_t^P = \left[x_t, y_t, \dot{x}_t, \dot{y}_t\right]^T$. This results in the linear state transformation matrices:

$$A^{(m^P_{\text{move}})} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, A^{(m^P_{\text{stop}})} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.1}$$

The vehicle observes pedestrian world positions $Y_t^P \in \mathbb{R}^2$ without velocities, resulting in the corresponding observation matrix $C^P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$.

For the context-based SLDS, the switching state M_t^P of the pedestrian motion model is encoded in the DBN as a categorical distribution $M_{t+1}^P = \text{Cat}\left(M_t^P, AL_{t+1}^P, HS_{t+1}^P, CC_{t+1}\right)$ as shown in Figure 6.2. The pedestrian awareness context S_t^P models whether the pedestrian sees the approaching vehicle. Head orientation HO_t^P forms the evidence. The context variable HS_t^P memorizes whether the pedestrian has seen the vehicle in the past, acting as a logical *OR* between previous HS_{t-1}^P and current S_t^P. The environment context AL_t^P models whether the pedestrian is near the curb, thus encoding where a pedestrian would normally stop to yield for oncoming traffic.

VEHICLE-RELATED CONTEXT STATES

The vehicle motion state is $X_t^V = [x_t, y_t, \dot{x}_t, \dot{y}_t]^T$. It uses a constant velocity model while driving, and a velocity decay model for braking:

$$A^{(m^V_{\text{move}})} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, A^{(m^V_{\text{stop}})} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & d & 0 \\ 0 & 0 & 0 & d \end{bmatrix} \tag{6.2}$$

The decay parameter $d = \sqrt[10]{0.5} \approx 0.93$ is empirically chosen to represent a velocity half-life of 0.5 s, i.e., the velocity becomes $d^{10} = 0.5$ of its initial value after 10 discrete time steps (0.5 s). This results in a mean initial deceleration of ~4.2 m/s^2 over the first second, thus reflecting moderate braking.Also, the vehicle V observes its own velocity, resulting in the observation matrix $C^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$.

For the vehicle, the context-based SLDS' switching state M^V is encoded as a categorical distribution $M^V_{t+1} = \text{Cat}\left(M^V_t, AL^V_{t+1}, HS^V_{t+1}, CC_{t+1}\right)$. The driver awareness context S^V_t models the driver's awareness of the pedestrian. It is inferred from the attention eccentricity HO^V_t, i.e., the absolute visual angle difference between the driver's center of gaze (or head direction) and the pedestrian. The context variable HS^V_t memorizes whether the driver has seen the pedestrian analogous to HS^P_t. The static environment context AL^V_t indicates whether the vehicle is at a distance from the pedestrian's crossing location where the driver can be expected to yield, assuming he/she has the intention to do so.

SHARED CONTEXT STATE

Both pedestrian and vehicle dynamics depend on CC_t, which indicates whether pedestrian and vehicle are on a collision course. It uses the minimum distance D^{min}_t obtained when linearly extrapolating the paths with their momentary estimated velocities [90].

6.2.2. INFERENCE

During inference the DBN states are propagated over time by incorporating observations in a forward filtering procedure (predict, update) following [61]. At each time step t, the entire state of the DBN is represented by the nine discrete latent states (four vehicle, four pedestrian, one shared) and two partially observable continuous latent states (X^V_t, X^P_t), see Figure 6.2. During the predict step, the value of each discrete latent state changes according to a fixed transition table, based on the values of its input states, i.e., each state's input nodes in Figure 6.2, including the state from the previous time step $t-1$ (dashed line). During the update step, observations are incorporated based on the context likelihood distributions, see Figure 6.3. The intermediate goal is to have the motion model switching states for both vehicle (M^V_t) and pedestrian (M^P_t) which represent the switching probability of the SLDS of each road user. The two continuous latent states X^V_t, X^P_t are propagated over time using observations (Y^V_t, Y^P_t) by standard Linear Dynamical

System (LDS) means, i.e., Kalman filter. Prediction into future without observation follows the same procedure, but without the update steps. Overall, this results in predicted motion states including uncertainties for both vehicle and pedestrian. To keep inference tractable, Assumed Density Filtering [77] is applied, resulting in the probability distributions of X_t^V, X_t^P to be each modeled by a Gaussian Mixture (K=2).

6.3. MODEL PARAMETER ESTIMATION

The DBN model parameters are set by performing a data-driven initialization step, followed by a gradient-based optimization step, using the dataset that is introduced in Section 6.4.

6.3.1. MODEL PARAMETER INITIALIZATION

Model parameters relate to motion dynamics and context. They are initialized similar to Kooij *et al.* [61].

MOTION DYNAMICS

The underlying motion models of M^V and M^P are represented by LDSes which model process noise Q and observation uncertainty R. Process noise Q of vehicle and pedestrian are set for both position and velocity states and are limited to diagonal matrix entries. Values were selected to reflect model uncertainty under typical velocity changes of drivers and pedestrians [82, 116]. Observation noise R is set to reflect typical variance of measurement noise for pedestrian detection and vehicle movement observed on-board the testing vehicle, see Section 6.4. The motion state transition matrices were obtained as follows. The vehicle motion state M^V was categorized as *braking* when such activity was detected, analogous to AL^V, and as *driving* otherwise. The pedestrian motion state M^P was categorized as *standing* in all scenarios where a pedestrian stops starting from three frames preceding time-to-event (TTE) = 0 (see Section 6.4.2 for definition of TTE), similarly to AL^P below. The motion state at all other time instants was categorized as *walking*. The motion state transition matrices were then obtained by counting and normalizing the occurrences of the respective transitions. The initial motion states assume the vehicle and pedestrian are driving and walking.

CONTEXT

To obtain the parameters for binary context states, their ground truth values need to be established; this is done in a two-step approach. In the first step, ground truth values were roughly obtained by setting some states to the same values for the entire scenario based on its definition (S^P, S^V, CC), by manual annotation ($AL^P = 1 \iff \text{TTE} = 0$), or by an automatic observable criterion ($AL^V = 1$ for all moments after first deceleration, i.e., pressing the brake pedal). This yields the context likelihood distributions as shown by the histograms in Figure 6.3. Parametric distributions were fitted by Maximum-Likelihood-Estimation and are shown by line plots. The parametric form of the distributions was chosen heuristically: Gaussian (DL^P, DL^V), Gamma (D^{min}, HO^V) or von-Mises distribution (HO^P).

In a second step, more accurate ground truth values for the context states were obtained on the basis of the obtained context likelihood distributions. For context states

AL^V, AL^P and CC, the values were re-assigned based on maximum likelihood criterion (e.g., $CC = 1 \iff D^{min} < 2.6$ m, see Figure 6.3a). For S^P and S^V, re-assignment was done heuristically. $S^P = 1 \iff HO^P \in [-30, 30]°$ was reassigned due to the largely overlapping distributions caused by miss-estimation of the head pose estimation algorithm. Similarly, $S^V = 1 \iff HO^V < 10°$ was reassigned whenever driver head orientation was used and $< 4°$ otherwise for the eye gaze orientation.

The transition matrices which represent the transition probabilities conditioned on the input states (i.e., incoming links in the DBN graph) were obtained by counting and normalizing the re-assigned binary context values between adjacent time steps. The transition probabilities of HS^V and HS^P are implemented as a binary OR in order to memorize the last state in accordance with their definition in Section 6.2.1.

The initial context states values were set conservatively at the beginning of each encounter: driver/pedestrian not looking, vehicle not near crossing location and pedestrian not at curb.

6.3.2. MODEL PARAMETER OPTIMIZATION

The gradient-based method of Pool *et al.* [93] was employed to obtain optimized model parameters. In short, the method performs back-propagation similar to neural networks on the DBN parameters on a differentiable loss function. The observation log likelihood of the vehicle and pedestrian under their respective predicted Gaussian distributions is optimized, see Eq. (6.4). All intermediate time-steps up to the prediction horizon are incorporated into the loss function to enforce a consistent path. Measurements with TTE $\in [-2.5\,\text{s}, 3.0\,\text{s}]$ are considered for optimization, to cover periods of typical motion dynamics. Missing intermediate measurements are ignored for optimization. TTE is defined in Section 6.4.

Optimization has been performed while enforcing properties of the DBN variables to keep the state representation interpretable, such as probabilities residing in $[0, 1]$ and process and observation noises remaining positive definite. The latter is also enforced to be diagonal matrices with variability along elements of main direction of travel to reduce degrees of freedom and obtain more stable convergence in the optimization process.

The model parameters chosen for optimization are: process noises (Q) of pedestrian and vehicle, transition probabilities, and context observation distribution parameters. The model was implemented in Python 3 using PyTorch 1.4 and was optimized using Adam [59].

6.4. DATASET

6.4.1. SCENARIOS

93 vehicle-pedestrian encounters with 4 trained drivers and 4 pedestrians were staged on two empty public roads. Each encounter consisted of a single pedestrian with the intention to cross the street in front of the approaching vehicle. The encounters represented nine disjoint scenarios (8-20 encounters each) with different combinations of situation criticality (collision course/sufficient time to cross), pedestrian behavior (stop at curb/cross), pedestrian awareness of the approaching vehicle (aware/unaware), vehicle behavior (brake/continue) and driver awareness of the approaching pedestrian

(aware/unaware). The included scenarios are listed in the left of Table 6.3.

All scenarios (except the anomalous scenario 9[a]) encode following behaviors:

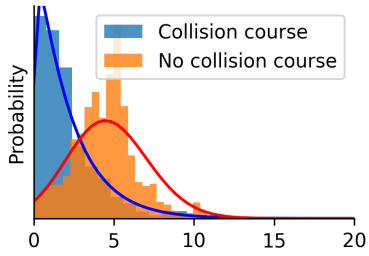
- An aware pedestrian will yield to the vehicle. Pedestrian awareness is inferred from pedestrian head pose.
- An aware driver brakes for an inattentive pedestrian approaching the curb. Awareness is inferred from driver head or gaze orientation.
- In non-collision-course crossing scenarios, both participants continue walking, respectively driving.
- Unaware participants continue walking/driving.

Scenarios 1 to 4 represent non-collision-course scenarios, meaning the pedestrian has sufficient time to cross. Scenarios 5 to 7 are safe through a change in behavior by either the driver or pedestrian due to awareness of the other participant. Scenario 8 represents a collision where both driver and pedestrian are unaware of each other's presence. Scenario 9[a] represents an anomalous scenario: the pedestrian crosses despite being aware of the approaching vehicle. The anomalous scenario is not considered for model parameter estimation.

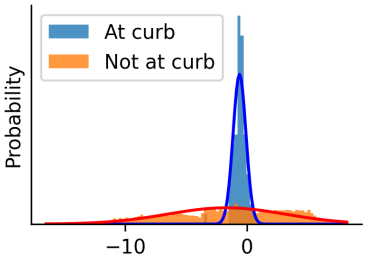
Pedestrians were instructed to either *"continuously observe the vehicle"* or to *"keep facing forward and don't look at the vehicle"*. Drivers were instructed to either *"keep looking at the pedestrian"* or to *"avoid looking at the pedestrian"* while approaching the pedestrian.

While scenarios 8 and 9[a] represent collisions, naturally, no actual collision took place during data collection. Instead, the vehicle was brought to a full stop before colliding with the pedestrian. The vehicle's velocity and position data were artificially replaced with a constant velocity model starting just before the onset of braking.

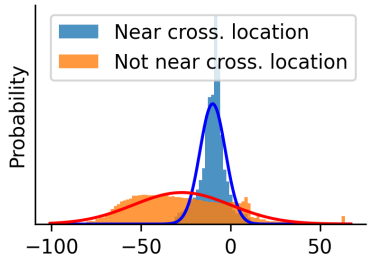
To ensure safety, the road was overseen to halt the experiments when other traffic entered the testing area. A co-driver provided verbal instructions on when to brake. Target driving speed was 20 km/h and pedestrians adopted their preferred walking pace.



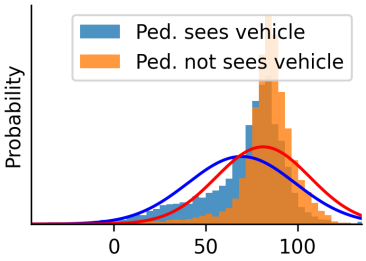
(a) D^{min}: Minimum distance along approach (m)



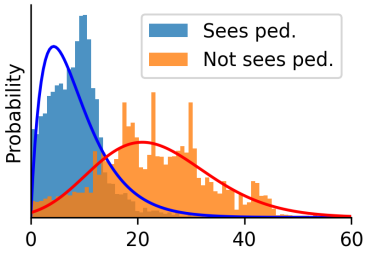
(b) DL^P: Pedestrian distance to curb (m)



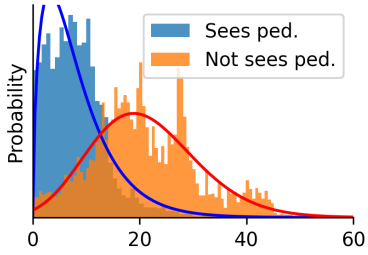
(c) DL^V: Vehicle distance to crossing location (m)



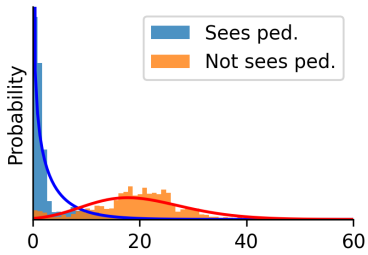
(d) HO^P: Pedestrian head orientation (deg)



(e) HO^V: Attention eccentricity (measured head pose) (deg)



(f) HO^V: Attention eccentricity (estimated head pose) (deg)



(g) HO^V: Attention eccentricity (estimated eye gaze) (deg)

Figure 6.3: Original and fitted context likelihood distributions. See Section 6.3.1 for details.

6.4.2. INSTRUMENTATION, MEASUREMENTS, AND GROUND TRUTH

All data were collected with a TU Delft experimental vehicle, whose instrumentation is described in further detail in [40]. Vehicle position, orientation and velocity are obtained from an ego-vehicle localization system which fuses differential GNSS, IMU, steering wheel angle and wheel ticks. This was implemented by the Robot Operating System (ROS) `robot_localization` package [79] and gained the transformations from vehicle frame to the world coordinate frame, which is set to identity at the start of the system. The GPS maintains a position accuracy of 4 cm while drift between GPS updates is limited to 0.8% per unit of distance traveled. The road was observed at 10 Hz using a forward-facing stereo camera (baseline 22 cm, 1936 × 1216 px) mounted behind the top-center of the windshield to obtain a dense stereo depth image of the scene in front of the vehicle.

Driver head pose and gaze were recorded with two systems. *Estimated* eye gaze and head pose were recorded with a high-end commercial off-the-shelf eye tracker (Smart-eye: 4-camera Smart Eye Pro dx 5.0, software version 8.2, running at 60 Hz with a gaze accuracy down to 0.5°). Secondly, *measured* head pose is obtained by a head-worn infrared-reflective marker tracked by an optical marker tracking system (Smarttrack) mounted on the rear seat head rest [107, 109]. Additionally, the driver was observed by a camera mounted above the speedometer for visual verification purposes. All sensor data were spatially calibrated and resampled to a target rate of 20 Hz.

Measured pedestrian positions on the ground plane were obtained in three successive steps:

1. 2D pedestrian bounding boxes were estimated from the forward facing camera by the Single-Shot-Multibox-Detector (SSD) of Braun *et al.* [12].

2. Distance to camera was found by median stereo disparity [53] of the 2D bounding box.

3. Transformation of this car-relative pedestrian position to ground plane positions in world coordinate frame was performed via ego-vehicle localization.

The time between the first pedestrian detection and the pedestrian reaching the curb was (min / max / mean = 1.3 s / 3.2 s / 2.9 s) over the various sequences. In that period, the pedestrian detection recall was 83 %.

Similarly to Kooij *et al.* [60], the pedestrian's focus-of-attention is inferred from pedestrian head orientation. The method of Braun *et al.* [13] is used to obtain a single yaw angle representing pedestrian head orientation.

In order to temporally compare prediction performance among the various scenarios, a semantically meaningful event was manually annotated for each sequence, as in [60], [58]. For scenarios where the pedestrian crosses, it represents the first frame where a pedestrian's foot crosses the curb. For scenarios where the pedestrian stops, it represents the moment where the last foot is placed on the ground near the curb. This implicitly defines time-to-event (TTE) for each time-step of each sequence (negative TTE: before event).

For each encounter, ground truth of the pedestrian position in the world coordinate frame is obtained. The pedestrian's target path of travel is defined in the world coordinate frame as a straight line and corresponds to the path the participants were instructed to

move along. The pedestrian ground plane location is then obtained by the intersection of the annotated path of travel with the vertical plane spanned by the image column of the hip point which was manually annotated in each frame. Map information and ego-vehicle localization is employed to estimate the location of the curb side.

6.5. EXPERIMENTS

To evaluate the incremental benefits of the DBN model components for an intelligent collision warning system, this section compares six models with varying access to the used context cues on their joint prediction performance of vehicle- and pedestrian-path and collision risk. Two evaluation metrics are adopted: the ability to predict driver and pedestrian location 1.5 s into the future, and collision risk across multiple prediction horizons. Evaluation is performed using 5-fold cross validation.

6.5.1. EVALUATION METRICS

For each time t, each model creates a predictive distribution $\widetilde{P}_{t\rightarrow t+t_p}(X_t)$ for state X_t and prediction horizon t_p. Based on the predictive distributions of both vehicle and pedestrian, individual path prediction performance and combined collision risk are evaluated.

PATH PREDICTION PERFORMANCE

Two performance metrics are used to evaluate path prediction performance [60] [58]: (a) Euclidean distance error between predicted expected position and future ground truth position GT_{t+t_p}:

$$\mathrm{error}\left(t_p|t\right) = \left|\mathbb{E}\left[\widetilde{P}_{t\rightarrow t+t_p}\left(X_t\right)\right] - \mathrm{GT}_{t+t_p}\right| \tag{6.3}$$

and (b) the log likelihood of the future ground truth position GT_{t+t_p} under the predictive distribution:

$$\mathrm{loglik}\left(t_p|t\right) = \log\left[\widetilde{P}_{t\rightarrow t+t_p}\left(\mathrm{GT}_{t+t_p}\right)\right] \tag{6.4}$$

loglik encapsulates both the spatial error and certainty about the position observation. Larger *loglik* values denote better prediction performance.

COLLISION RISK

The probability for a collision is determined by taking the integral of the predictive distributions over a collision area, which is defined by all possible intersections between vehicle and pedestrian locations. Let $\widetilde{P}_{t\rightarrow t+t_p}(X_t) = \mathcal{N}(\mu_{t\rightarrow t+t_p}, \sigma^2_{t\rightarrow t+t_p})$ bea single Gaussian predictive position of either pedestrian P or vehicle V. The *combined* predictive position is then defined as $\widetilde{P}^{\phi}_{t\rightarrow t+t_p}(X^{\mathrm{P}}_t, X^{\mathrm{V}}_t) = \mathcal{N}(\mu^P_{t\rightarrow t+t_p} - \mu^V_{t\rightarrow t+t_p}, (\sigma^P_{t\rightarrow t+t_p})^2 + (\sigma^V_{t\rightarrow t+t_p})^2)$. The collision risk predicted from t for $t+t_p$ is given by:

$$\mathrm{CR}\left(t_p|t\right) = \int_{A^{\phi}} \widetilde{P}^{\phi}_{t\rightarrow t+t_p}(X^{\mathrm{P}}_t, X^{\mathrm{V}}_t)\, dX^{\mathrm{P}}_t\, dX^{\mathrm{V}}_t \tag{6.5}$$

with A^{ϕ} being the combined spatial spatial extent of vehicle and pedestrian. If the predictive distributions for the vehicle and the pedestrian are represented as Gaussian

Mixtures (SLDS and DBN variants), the overall collision risk is given by the weighted pairwise collision risk between the Gaussian Mixture components. This extends the collision risk estimation method of Braeuchle *et al.* [10].

For the application of collision risk warning, collision probability has to be classified into collision or no collision, and classification performance requires a ground truth for collision outcome. Collision ground truth is defined as true for any time instance where the vehicle and pedestrian ground truth overlap given their position and spatial spatial extent. In order to assess the collision risk prediction performance at various prediction horizons, a fixed false positive rate (FPR) is selected and the attainable true positive rate (TPR) is found for each prediction horizon t_p.

6.5.2. MODEL VARIANTS

This chapter evaluates four context-aware models, including the method of Kooij *et al.* [61], which differ in their access to pedestrian and vehicle context, and compare them to two context-agnostic models. An overview of the used context cues of the models is given in Table 6.2. All models were optimized individually as described in Section 6.3.

CONTEXT-AGNOSTIC *LDS*

Both linear dynamical systems for pedestrian and vehicle path prediction are instantiated by constant velocity motion models.

CONTEXT-AGNOSTIC *SLDS*

Vehicle and pedestrian motion are both modeled by context-agnostic SLDSes with the same underlying motion models as the context-aware models (driving/braking, walking/standing) described below.

CONTEXT-AWARE MODELS WITH VARYING PEDESTRIAN- AND VEHICLE-CONTEXT

This section analyzes four variants of the model presented in Figure 6.2 which take different amounts of context into account: *DBN.p* represents the context-based pedestrian path prediction method of Kooij *et al.* [61]. The method is **p**edestrian-aware, vehicle-agnostic

Table 6.2: Context cues and number of motion models per road user used in the models. DBN suffixes denote used context: p: pedestrian [61]; v: vehicle (AL^V); h: driver head pose; g: driver gaze. E.g., *DBN.pvg* uses pedestrian, vehicle and driver eye gaze awareness context.

Context cue	LDS	SLDS	DBN.p [61]	DBN.pv	DBN.pvh	DBN.pvg
Pedestrian at-curb	-	-	x	x	x	x
Pedestrian awareness	-	-	x	x	x	x
Collision course	-	-	x	x	x	x
Vehicle near-crossing	-	-	-	x	x	x
Driver awareness	-	-	-	-	head pose	eye gaze
# Ped. motion models	1	2	2	2	2	2
# Veh. motion models	1	2	2	2	2	2

and driver-agnostic. It models the vehicle dynamics as a context-agnostic SLDS. *DBN.pv* is **p**edestrian, **v**ehicle-aware and extends *DBN.p* with vehicle static environment cues but remains driver-agnostic. It includes proximity of the vehicle to the crossing location of the pedestrian (AL^V). *DBN.pvh* additionally uses driver **h**ead pose as an awareness cue (S^V). *DBN.pvg* uses driver eye **g**aze instead of driver head pose.

6.5.3. PATH PREDICTION

Table 6.3 depicts average path prediction performance over various encounters of a certain scenario in terms of *loglik* and Euclidean distance error of both pedestrian and vehicle for a prediction horizon $t_p = 1.5\,\text{s}$ averaged over periods where typical changes in dynamics occur (pedestrian: TTE $\in [-0.5\,\text{s}, 2.0\,\text{s}]$, vehicle: TTE $\in [-0.5\,\text{s}, 3.0\,\text{s}]$; TTE ranges define times where predictions are made for). Let us consider three scenario types.

Table 6.3: Scenario decomposition (left), mean path prediction performance in terms of *loglik* (center) and Euclidean distance error (right) of various models for a prediction horizon of $t_p = 1.5\,\text{s}$. The top and lower halves of the table capture the prediction performances of pedestrian and vehicle along the dimension of main travel (i.e. lateral and longitudinal vs. vehicle main axis). See Section 6.5.2 for model definitions. Higher *loglik* and lower Euclidean distance error denote better prediction performance. Bold numbers denote best-performing model per scenario. Grey rows denote scenarios with a change in dynamics of the respective road user.

Scen.	CC	Ped. stops	Ped. sees	Veh. stops	Driver sees	LDS	SLDS	DBN p [61]	DBN pv	DBN pvh	DBN pvg	LDS	SLDS	DBN p [61]	DBN pv	DBN pvh	DBN pvg
						Pedestrian 1.5 s *loglik*						Pedestrian 1.5 s Euclidean error (cm)					
1	0	0	0	0	0	−3.3	−2.2	**−2.1**	**−2.1**	−2.2	−2.2	64	99	**48**	51	52	51
2	0	0	0	0	1	−2.8	−2.7	−2.5	−2.5	**−2.4**	**−2.4**	**83**	140	112	110	110	111
3	0	0	1	0	0	−9.2	−3.5	**−3.1**	**−3.1**	−3.6	−3.7	77	133	**68**	71	77	73
4	0	0	1	0	1	−9.0	−2.3	−2.3	**−2.2**	−2.3	−2.3	54	73	55	50	**46**	49
5	1	1	1	0	1	−4.0	−2.4	**−1.8**	**−1.8**	−2.2	−2.2	122	131	**84**	86	91	91
6	1	1	1	0	0	−4.2	−2.5	**−1.7**	**−1.7**	−1.8	−1.8	114	131	**83**	87	87	87
7	1	0	0	1	1	**−1.1**	−1.5	−1.9	−1.8	−1.7	−1.7	**58**	90	71	70	70	70
8	1	0	0	0	0	**−1.0**	−1.3	−2.0	−1.9	−1.9	−1.9	**52**	74	63	61	63	63
9[a]	1	0	1	0	0	**−1.5**	−1.8	−2.1	−2.0	−2.0	−2.0	**63**	100	79	77	73	73
non-anomalous, motion change (5-6)						−4.1	−2.5	**−1.8**	**−1.8**	−2.0	−2.0	118	131	**84**	87	89	89
non-anomalous, no motion change (1-4, 7-8)						−4.4	**−2.3**	**−2.3**	**−2.3**	−2.4	−2.4	**65**	102	70	69	70	70
						Vehicle 1.5 s *loglik*						Vehicle 1.5 s Euclidean error (cm)					
1	0	0	0	0	0	−6.2	**−2.2**	−2.8	−2.8	−2.8	−2.8	54	53	**46**	52	55	55
2	0	0	0	0	1	−38.0	−7.4	−8.8	**−6.0**	−6.1	−6.1	60	62	**49**	53	55	55
3	0	0	1	0	0	−31.2	**−6.1**	−7.9	−7.9	−7.0	−7.0	48	52	**39**	44	51	50
4	0	0	1	0	1	−12.9	**−2.8**	−3.7	−3.6	−3.7	−3.8	63	66	**55**	56	58	58
5	1	1	1	0	1	−4.5	**−1.5**	−2.4	−2.1	−2.0	−2.0	**48**	54	**48**	117	69	69
6	1	1	1	0	0	−3.4	**−1.4**	−2.0	−2.0	−1.8	−1.8	43	52	**40**	103	61	61
7	1	0	0	1	1	−7.8	−2.7	−2.6	**−2.1**	−2.2	−2.2	245	189	195	**149**	175	175
8	1	0	0	0	0	**−1.0**	**−1.0**	−1.6	−1.7	−1.6	−1.6	46	47	**39**	81	45	45
9[a]	1	0	1	0	0	**−1.1**	**−1.1**	−1.6	−1.8	−1.7	−1.7	38	47	**34**	78	45	45
non-anomalous, motion change (7)						−7.8	−2.7	−2.6	**−2.1**	−2.2	−2.2	245	189	195	**149**	175	175
non-anomalous, no motion change (1-6, 8)						−13.9	**−3.2**	−4.2	−3.7	−3.6	−3.6	52	55	**45**	72	56	56

NORMAL SCENARIOS WITH NO MOTION CHANGE

First the normal scenarios are considered where no motion change occurs for a certain road user (i.e. scenarios 1-4 and 7-8 for the pedestrian, and scenarios 1-6 and 8 for the vehicle; the respective average performances are listed in two separate rows of Table 6.3).

It can be seen that the *LDS* for that road user has a comparatively poor *loglik* overall (−4.4 and −13.9, resp.), as the uncertainty region of its single-Gaussian state representation is large to account for possible motion changes. On the other hand, its maximum likelihood estimate is comparatively accurate: the Euclidean distance error is smaller than that of other models (65 cm and 52 cm, for pedestrian and vehicle resp.); this is to be expected as its linear model precisely fits the actual motion.

It can also be observed that context-aware models are at least on-par-with their context-agnostic (multi-motion) counterparts; cases of outperformance suggest that the context in the former provides more selective guidance when a motion change is probable. Specifically, models that incorporate pedestrian context (all DBN variants) are on-par-with (outperform) *SLDS* in terms of the *loglik* (Euclidean distance error) metric for the pedestrian. Models that incorporate vehicle context (*DBN.pv*, *DBN.pvh* and *DBN.pvg*) are on-par-with *SLDS* in terms of the *loglik* and Euclidean distance error metric for the vehicle.

NORMAL SCENARIOS WITH MOTION CHANGE

Now the normal scenarios are considered where motion change occurs for a certain road user (i.e. scenarios 5-6 for the pedestrian, and scenario 7 for the vehicle; the respective average performances are listed in two separate rows of Table 6.3).

It can be seen that the context-aware models for a road user mostly outperform their context-agnostic counterparts (*LDS* and *SLDS*) in terms of *loglik* and Euclidean distance error for that road user. There is an observation that having the full context of a road user does not necessarily improve performance for that road user as opposed to using only partial context (e.g. for the vehicle, *DBN.pvh* and *DBN.pvg* underperform *DBN.pv* on Euclidean distance error.)

Adding context related to the other user does not improve performance for the original road user (e.g. adding vehicle context *DBN.pvh* and *DBN.pvg* does not outperform pedestrian prediction performance by *DBN.p*). An outperformance might have been expected, as a motion change indicates an interaction between the road users, where such other road user context could be helpful. Apparently, the motion coupling by means of the *CC* state variable in the DBN is (too) weak, and is possibly overshadowed by data issues (e.g. measurement noise, insufficient data).

Figure 6.4 shows a temporal analysis of vehicle path prediction performance for sequences where the vehicle stops (scenario 7). While the vehicle approaches the pedestrian with constant velocity (TTE < -0.2 s), the three compared models (*LDS*, *SLDS*, *DBN.pvg*) show similar performance. As the vehicle slows down, both *LDS* and *SLDS* increase in spread over various sequences (shown by the standard deviations) and gradually decrease in vehicle *loglik*. The *SLDS* model adapts more quickly to the change of dynamics (switch from driving to braking) compared to the *LDS*. The *DBN.pvg* model variant anticipates the change in motion dynamics resulting in a higher *loglik* and less uncertainty than the context-agnostic models, therefore resulting in a better path prediction performance for the vehicle.

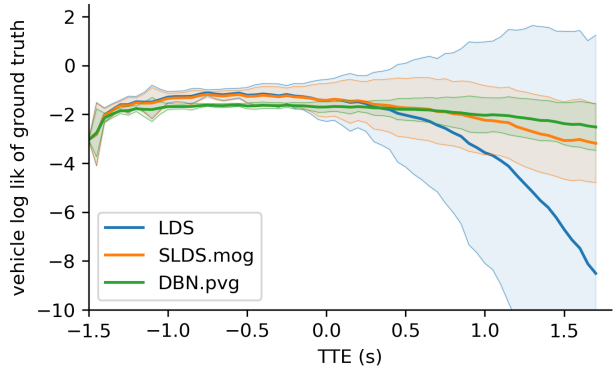


Figure 6.4: *Loglik* and standard deviation over time for a braking vehicle (scenario 7) for a prediction horizon $t_p = 1.5$ s, and drawn at the moment for which the prediction was created (i.e., the values shown at TTE = 0.0 s were predicted from measurements of TTE = −1.5 s). The vehicle initiates braking for the crossing pedestrian between −1.8 s and 0.6 s, with most vehicles braking from 0.0 s onward.

ANOMALOUS SCENARIO

Finally, let us consider the anomalous scenario 9[a]. It is anomalous as the pedestrian crosses despite seeing the vehicle. Table 6.3 shows a lower prediction performance of the context-aware models (all *DBN* variants) regarding the pedestrian compared to the context-agnostic models (*SLDS* and *LDS*). This is no surprise, as the context-aware models were trained to expect stopping behaviour. Despite this, performance degrades gracefully, since the measurements of the walking pedestrian allow the context-aware models to infer decent motion state estimates.

Figure 6.5 shows a comparison between driver gaze (*DBN.pvg*) and driver head pose (*DBN.pvh*) as contextual cue for S^V (sees-pedestrian). For $S^V = 1$, driver gaze provides higher classification confidence in HS^V (has-seen-pedestrian) compared to head pose. For $S^V = 0$, both models incorrectly believe that the driver has seen the pedestrian for a similar fraction of sequences. However, this classification accuracy did not yield a better vehicle path prediction performance when comparing *DBN.pvg* to *DBN.pvh* in Table 6.3. This can be attributed to the memorizing effect of HS^V.

Measured driver head pose (Smarttrack) provided virtually identical results to estimated head pose (Smarteye) on all scenarios, and was therefor excluded from analysis.

6.5.4. COLLISION RISK ESTIMATION

This section first compares how collision risk estimates evolve over time for the *LDS*, *SLDS* and *DBN.pvg* models on two exemplary sequences with changing vehicle dynamics (scenario 7) and collision (scenario 8), followed by an assessment of overall collision risk prediction performance as function of prediction horizon.

SCENARIO-BASED COLLISION RISK

Figure 6.6a shows collision risk prediction for a sequence from scenario 7, where the vehicle brakes due to an aware driver. Thus, a low predicted collision risk is expected.

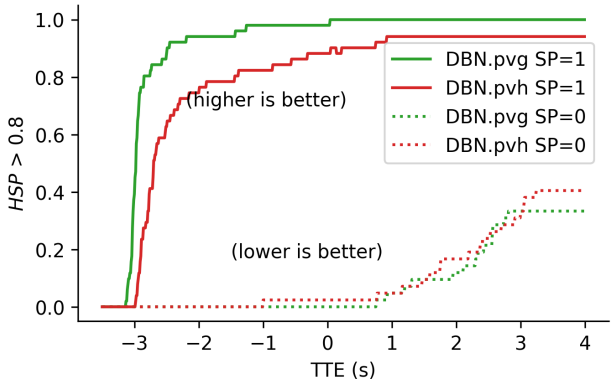
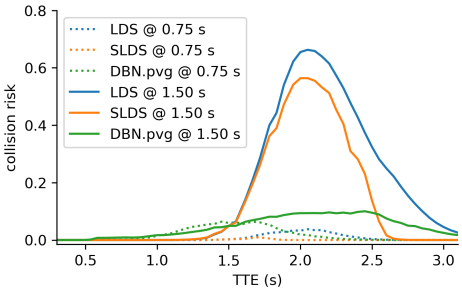
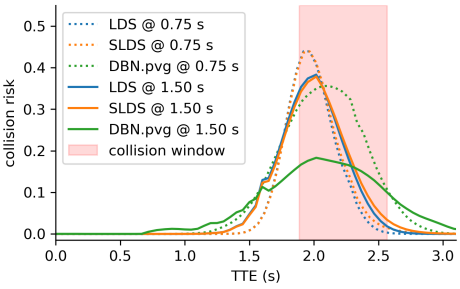


Figure 6.5: Classification performance of *DBN.pvg* and *DBN.pvh* on the hidden HS^V state on sequences where driver is instructed to be attentive ($S^V = 1$) and inattentive ($S^V = 0$).

For a prediction horizon $t_p = 0.75\,\text{s}$, all models predict a negligible collision risk (dashed lines). Predicting $t_p = 1.5\,\text{s}$ into future, the *LDS* and *SLDS* models anticipate a collision risk of 66% and 56% respectively while the *DBN.pvg* model keeps a collision risk below 10% throughout the sequence.



(a) Sequence from scenario 7. Lower collision risk denotes better performance.



(b) Sequence from collision scenario 8. Higher collision risk denotes better performance. The collision window CW is shaded in red.

Figure 6.6: Collision risk estimates obtained from different models for a braking vehicle (a) and collision (b) sequence. TTE indicates the time for which the predictions were made. Values are shown for prediction horizons t_p of 0.75 s and 1.5 s.

Figure 6.6b shows collision risk over time for one sequence from the collision scenario (scenario 8), where both the vehicle and the pedestrian continue their respective motion, being unaware of each other. The *collision window* depicts all time instances defined as a collision in accordance with Section 6.5.1, i.e., where the geometries of vehicle and pedestrian overlap. Predicting 0.75 s into the future, all compared models (*LDS*, *SLDS*, *DBN.pvg*) depict similar maxima of collision risk within the collision window. With

increasing prediction horizon, each model becomes less certain, resulting in a lower predicted collision risk value.

The maxima are above 18% within the collision window for the exemplaricly depicted sequence. Figure 6.6 further shows that only for *DBN.pvg*, there exists a range of collision risk thresholds (10%–18%) for which a collision warning is triggered in the collision sequence (Figure 6.6b) but not in the non-collision sequence (Figure 6.6a).

OVERALL COLLISION RISK PREDICTION

To examine how collision risk prediction performance changes with prediction horizon t_p, a FPR of 1% is selected and the attainable TPR as a function of t_p is evaluated, see Figure 6.7. One observes that the context-agnostic models (*LDS* and *SLDS*) significantly under-perform the context-aware models (*DBN* variants). For a prediction horizon up to 0.75 s, all DBN variants achieve a TPR close to 1.0. They continue to perform similarly until a prediction horizon of about 1.3 s, after which point the driver-aware models *DBN.pvh* and *DBN.pvg* obtain a small edge. Towards a horizon of 2.0 s, the TPR of the models drops towards 10%.

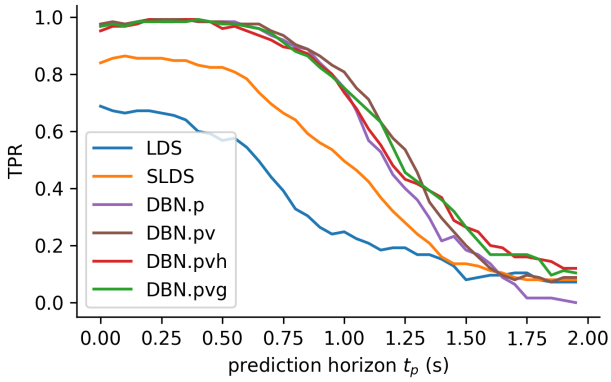


Figure 6.7: Collision risk TPR of different models obtained under a 1% FPR for various prediction horizons. Higher values denote better performance.

6.6. DISCUSSION

Path prediction performance was evaluated in three scenario types within a time interval of a few seconds around a potential motion change: in normal scenarios with no motion change, in normal scenarios with motion change and in an anomalous scenario. Reporting aggregate performance would not have been very insightful. This is because in reality, the time steps in which "normal" scenarios apply with no motion changes vastly outnumber the two other scenario types. Just considering aggregate performance would strongly favor simple models like the *LDS* (or a parameter setting of a more complex model that essentially implements such a simple model). However, the time instants involving motion changes should arguably carry more weight, as they might strongly induce changes in collision risk. Listing separate performance values for various scenario types allows to side-step this weighting issue for now.

For normal scenarios with no motion change, the single-motion model *LDS* performs best in terms of Euclidean distance error, albeit with by far the worst *loglik* performance of all models. Context-aware models (*DBN.p*, *DBN.pv*, *DBN.pvh*, *DBN.pvg)* were at least on-par-with their context-agnostic (multi-motion) versions (*SLDS*). They remained competitive with the *LDS* on Euclidean distance error. The normal scenarios with motion changes are those settings where the context-aware models can potentially shine. Indeed, the context-aware models were found to mostly outperform their context-agnostic counterparts (*LDS* and *SLDS*). Anomalous situations which defy the anticipated motions, but still occur in real-world traffic, provide a challenge to a context-aware model. They might contradict the expert knowledge encoded in the DBN structure or will not adhere to the parameters estimated on a training set. Fortunately, the probabilistic modeling allows for softer decisions: the switch of motion dynamics not only depends on the pre-conditioning context, but also on the current positional observations. Indeed, the performances of context-aware models were shown to remain competitive with that of context-agnostic counterparts.

Overall, one observes that the models using both pedestrian and vehicle context (*DBN.pv*, *DBN.pvg*, *DBN.pvh*) performed best over the three time scenario types. Full context was not shown to improve path prediction performance (i.e. *DBN.pvg* and *DBN.pvh* not outperforming *DBN.pv*). While *DBN.pv*, *DBN.pvh* and *DBN.pvg* encode typical vehicle braking locations, variation in braking behavior seems to limit the predictive value of the driver awareness cue. Contrary to the initial expectations, measuring driver gaze (*DBN.pvg*) yielded similar path prediction and collision risk estimation performance compared to measuring driver head pose (*DBN.pvh*), i.e. see Figure 6.7.

However, when multiple road users or driving distractions are introduced, it is likely that driver awareness will be dis-ambiguated more accurately from gaze compared to head-pose. Other fixation-related metrics may provide further insights in driver awareness, such as number of fixations, total fixation duration and angle of first saccade landing within 2° of the pedestrian [128], though such evaluations would require natural as opposed to instructed viewing behavior, and other spatial regions competing for attention.

In this chapter, mutual awareness and interaction between vehicle and pedestrian were chosen to be modeled loosely, by means of the shared context state *CC* (collision course) of the respective DBN sub-graphs. This has the advantage that it could easily scale-up to multiple road users, as their DBN sub-graphs can be designed and optimized individually, and the number of dependencies grow linearly. On the other hand, some limitations result from this loose motion coupling. The driver-aware models (*DBN.pvh, DBN.pvg*) encode the following: if one road user *A* is aware of the other *B*, this influences the motion of *A* which affects the shared collision course latent state *CC*, which in turn influences the motion of *B*. Not modeling the dependency between awareness of *A* and motion of *B* directly might lead to decreased performance. Consider the path prediction performance of the vehicle in scenarios 5 and 7. In both scenarios, the driver sees the pedestrian, however, only in scenario 7 the vehicle stops (due to the unaware pedestrian). The fact that the vehicle motion in the driver-aware models is not directly influenced by the pedestrian's awareness might contribute to why *DBN.pvh* and *DBN.pvg* are not the best performing models for scenario 7.

7

Conclusion and Future Work

This thesis addressed the perception of driver and pedestrian to extract mutual awareness for path prediction in the domain of intelligent vehicles. It has contributed to multiple components along the processing chain of an intelligent vehicle. Firstly, a large, naturalistic driver head pose dataset has been created as a foundation to develop a head pose estimation method, which estimates continuous, 6 degrees of freedom (DOF) head pose from single camera images. Perceiving the outside environment of an intelligent vehicle, this thesis has presented a person detection system that estimates 3D location, spatial extent (i.e., length, width and height), and yaw orientation of pedestrians around the vehicle from camera and lidar. Finally, driver head pose, vehicle ego-motion, and pedestrian features (like location, body and head orientation) have been used to develop a probabilistic path prediction system that takes into account the mutual awareness of driver and pedestrian, therefore allowing for an improved, probabilistic prediction of ego-vehicle and pedestrian paths in scenarios where a pedestrian might or might not cross the road in front of the approaching vehicle.

7.1. Conclusion

This section presents the main findings and draws conclusions along the methodical chapters.

Driver Head Pose Benchmark

Chapter 3 introduced *DD-Pose*, a large-scale driver head pose benchmark featuring multi-camera images of 27 drivers captured during 12 naturalistic driving scenarios. The benchmark contains 330 k frames with high-resolution stereo images from a driver camera, accompanied by an interior camera and driving metadata such as velocity and yaw rate. It provides per-frame head pose measurements and occlusion annotations. Precise head pose is measured by a novel calibration device. All sensors have been calibrated extrinsically and intrinsically, and are synchronized.

Compared to bounding box annotations, which are defined in image-space, 6 DOF head pose ground truth has been shown to be more difficult to obtain. An invasive, head-

worn measurement device has been employed to get accurate ground truth (as opposed to head pose "measured" in the camera image or from depth data). While the measurement device needs to be calibrated once per subject (<1 minute), all successive head pose annotations come with no additional manual effort. Diversity by a large number of training subjects (i.e., different drivers) is more difficult to obtain in contrast to collecting pedestrian images in urban scenes. Subjects have to drive the vehicle – and in the scope of a research vehicle – also require permission to drive the vehicle.

Compared to previous datasets, *DD-Pose* depicts a broader distribution of head poses at a higher resolution while providing absolute 6 DOF head pose. The experiments showed that *DD-Pose* provides challenges for a current State-of-Art (SoA) method due to its richness in extreme non-frontal head poses. At the time of writing, 58 researchers from 23 different countries have registered for the use of the dataset. It has proven valuable for the development of the head pose estimation method presented in Chapter 4 which will be discussed in the next section.

Head Pose Estimation Leveraging Camera Intrinsics

Chapter 4 has tackled the problem of 6 DOF head pose estimation from single camera images and their associated camera intrinsics in the domain of driver observation. This domain poses interesting in-car applications and challenges such as difficult illumination conditions and large out-of-plane rotations.

It was shown that explicit use of camera intrinsics is required for precise head pose estimation and it is used consistently within the presented novel intrinsics-aware head pose estimation method.

For decades, Euler angles or Quaternions have dominated the rotation representation within head pose estimation methods. This has led to several drawbacks (such as gimbal-lock, discontinuities, normalization, and ambiguities) which were mitigated by more complex models. The method presented in Chapter 4 employs a continuous rotation representation (SVDO$^+$) which simplifies the network architecture to a simple regression head and a pose conversion which yields a rotation in SO(3), the special orthogonal group spanning all rotations in 3D space.

Evaluations on the challenging in-car dataset *DD-Pose* from Chapter 3 have shown that leveraging camera-intrinsics alongside a continuous rotation representation results in a balanced mean absolute error (BMAE) of 5.8° compared to the intrinsics-agnostic SoA baseline (14.8°). Also, using an unbiasing data sampling strategy lowered the BMAE on the *hard* subset (extreme non-frontal rotations and occlusions) from 15.3° to 9.5°. The proposed method showed translation errors of 22/29/41 mm over the *easy*/ *moderate*/ *hard* subsets in the *DD-Pose* test set.

Overall, intrinsics-awareness and using a continuous rotation representation allowed for a simple architecture that yields robust head pose estimates across a broad spectrum of head poses. Furthermore, a runtime of <20 ms makes in-vehicle deployment possible.

Deep End-to-end 3D Person Detection from Camera and Lidar

Chapter 5 presented a novel deep end-to-end method for 3D person detection from camera images and lidar point clouds. The method does not rely on hand-crafted features. Instead, it learns high-level features from both camera images and lidar point clouds. Point cloud features are extracted using voxel feature encoders. Experiments on the KITTI

3D object detection benchmark show that the presented method outperforms the prior SoA by 2.2 percentage points with an average precision of 47.1% on moderate difficulty (validation set).

The method described in Chapter 5 is an early adopter of the end-to-end scheme training of multi-modal data (camera and lidar). Since then, the SoA has advanced in the research area of end-to-end training, by extending the differentiable path along the functional chain from detection via fusion up to motion planning [142], but also by extending in terms of fusing multiple cameras [71, 92].

DRIVER AND PEDESTRIAN MUTUAL AWARENESS FOR PATH PREDICTION

Chapter 6 presented a novel method for vehicle-pedestrian path prediction that takes into account the awareness of the driver and the pedestrian towards each other. The method jointly modeled the paths of a vehicle and a pedestrian within a single Dynamic Bayesian Network (DBN). Subsequently, collision risk was estimated by a probabilistic intersection operation. Overall, this work demonstrated an integrated system from on-board sensing up to collision warning.

Incremental benefits of pedestrian- and vehicle-context in six models with varying access to the used context cues were evaluated, namely Linear Dynamical System (*LDS*, one motion model), Switching Linear Dynamical System (*SLDS*, two motion models), *DBN.p* (**p**edestrian-aware), *DBN.pv* (**p**edestrian-aware, **v**ehicle-aware and driver-agnostic), *DBN.pvg* (driver **g**aze as awareness cue) and *DBN.pvh* (driver **h**ead pose as awareness cue). For validation, real-world data obtained by on-board vehicle sensing (stereo vision, GNSS, and proprioceptive) were used and consisted of vehicle and pedestrian encounters, spanning various awareness conditions and dynamic characteristics of the participants. For normal scenarios with no motion change, the single-motion model *LDS* performed best in terms of Euclidean distance error, albeit with the worst *loglik* (log-likelihood of ground truth position within predicted distribution) by far of all models. Context-aware models (*DBN.p*, *DBN.pv*, *DBN.pvh*, *DBN.pvg)* were at least on-par-with their context-agnostic (multi-motion) versions (*SLDS*). They remained competitive with the *LDS* on Euclidean distance error. In the normal scenarios with motion changes the context-aware models were found to mostly outperform their context-agnostic counterparts (*LDS* and *SLDS*). Even in an anomalous scenario, the performances of context-aware models were shown to remain competitive with that of context-agnostic counterparts. Overall, models using both pedestrian and vehicle context (*DBN.pv*, *DBN.pvg*, *DBN.pvh*) performed best on path prediction. This was also reflected in collision risk estimation performance. For example, the collision risk warning true positive rate (TPR) was raised from 18% (pedestrian-aware model *DBN.p* of Kooij *et al.* [61]) to 27% for *DBN.pvg* for a prediction horizon of 1.5 s and a false positive rate (FPR) of 1% over the dataset.

One of the main insights of Chapter 6 is that context cues can help to improve path prediction. However, simply using more complex motion models with additional context cues does not necessarily help prediction performance, if those context cues are not sufficiently informative or they cannot be reliably inferred from sensor measurements. Differences in path prediction performance between context cues can be subtle and might also not materialize due to small data sample effects and due to errors in the estimation of ground truth.

DBNs provide a versatile structure to model expert knowledge. E.g., in the presented DBN, it is predefined what type of motion models are used per traffic participant, and which observations influence which latent states. Similarly, awareness of the other traffic participant is represented by "having turned one's head towards the other traffic participant". This faces two issues. First, having turned a head in a certain direction does not necessarily reflect the focus of attention. Second, there is no "forgetting of awareness", i.e., the model assumes awareness persists over the duration of the encounter. Encoding expert knowledge implicitly encodes further assumptions. One assumption the model has is that there is a single pedestrian in the ego-vehicle's environment, and that the model has knowledge of a potential location of crossing, such as a zebra crossing or a crossing hot-spot, potentially obtainable by a map. More assumptions are implicitly represented by the dataset which is used for training the model, but also for validation. The dataset used in Chapter 6 covers multiple combinations of the vehicle braking/not braking, the driver looking/not looking, the pedestrian stopping/not stopping and the pedestrian looking/not looking. Yet this only covers a subset of the complex traffic interactions that can occur. The complex interactions need to be covered representatively in the dataset. At some point modeling interactions with a DBN structure reaches limitations and purely data-driven approaches need to be employed.

OVERALL OBJECTIVE

The overall goal of this thesis is a safer navigation of traffic with driver-pedestrian encounters. Towards this goal, this thesis presented a framework for joint path prediction of ego-vehicle and a pedestrian working on data of on-board sensors of a vehicle. The main components working together are: a camera-based driver head pose estimation component (looking-in; Chapter 4), a 3D pedestrian detection component (looking-out; Chapter 5), and a path prediction component combining the output obtainable by the two latter components over time (Chapter 6). Technical challenges in integrating the sensors into research vehicles have been overcome, such as calibration of sensors of different modalities inside and outside the vehicle, time synchronization and recording. The components can be integrated in a common framework, yet they have been developed and evaluated independently within the scope of this thesis. While it makes sense to optimize each component in isolation to obtain optimal outputs for their task, what eventually matters more is the output performance of the overall system. The evaluations of Chapter 6 have shown that fine-grained driver eye gaze performed equally good in terms of path prediction compared to coarse head pose. Admittedly, this might be due to limitations of the path prediction component.

What is the gap to close when integrating the framework into an Advanced Driver-Assistance System (ADAS) of a series-vehicle? The key performance indicators measuring the performance of the overall system would be the number of true positive emergency brakes (or warnings) for pedestrian crossing encounters, and the number of false positive emergency brakes for non-crossing encounters on a large, complex, representative validation dataset. While those target numbers are not clear to define, vehicle manufacturers tune ADAS in an SAE level 0–2 system towards a low number of false positive brakes for customer acceptance reasons. This is possible because the driver is still responsible at those SAE levels. Leaving the driver out of the picture for once, the Euro New Car

Assessment Program (NCAP) have defined a catalog of AEB VRU tests to perform on crossing pedestrian dummies [35]. These reflect important, yet isolated scenarios. E.g. *Car-to-Pedestrian Nearside Adult 25% (CPNA-25)* is defined as follows: "a collision in which a vehicle travels forwards towards an adult pedestrian crossing its path walking from the nearside and the frontal structure of the vehicle strikes the pedestrian at 25% of the vehicle's width when no braking action is applied." The pedestrian dummy speed is fixed to a static direction and constant velocity of 5 km/h. The datasets and experiments presented in this thesis show this setting does not reflect the complex movement pedestrians exhibit in real-life and therefore defines a lower bound for series vehicles. For now, customer acceptance puts a higher requirement on current systems, e.g., a vehicle known to slowly navigate urban traffic with pedestrians present on the sidewalk will find less acceptance.

The components presented in this thesis can be deployed into an ADAS with little adaptation, and improved further along the road. Driver cameras have found their way into the series market. Cameras facing the outside world are already part of series vehicles with an NCAP five star rating. The method of Chapter 5 relies on a lidar sensor besides the camera. It can be replaced by a camera-only pedestrian detection system. The presented methods run in real-time and can be deployed onto in-vehicle computers. The traffic scenario covered in this thesis is limited to a single pedestrian potentially crossing the road. Despite its current limitations the presented collision warning system would already improve road safety and can be extended to more complex traffic situations in future. Potential performance gaps are to be closed by representative data of the inside and the outside of the vehicle.

7.2. FUTURE WORK

This section discusses potential improvements in driver observation, environment perception, and driver-road-user interaction for path prediction. It concludes with the importance of data and a recommendation on how to use and share data to accelerate the development of automated driving.

DRIVER OBSERVATION

This thesis has demonstrated a robust method for driver head pose estimation in Chapter 4. Besides an accurate estimate, a component that consumes head pose will also benefit from the (un)certainty of the head pose estimate. To that end, uncertainty can be integrated into the neural network and trained alongside the head pose estimate. As the rotation of the presented method is represented in SO(3), the special orthogonal group spanning 3D rotations, representation of rotation uncertainty via Bingham belief [91] would be a natural choice.

The single-frame method has a recall of 93% on the *hard* subset which is not sufficient for safety-relevant applications. A driver observation system would integrate the 6 DOF estimates in a temporal filtering scheme to obtain more robust pose estimates and increase the recall. The presented head pose estimation method employs a two-stage detection network and relies on bounding box proposals (anchors). Aspect ratio and size of bounding box proposals have already been optimized for human heads. Still, further work could experiment with adapting one-stage detection networks which skip

the bounding box proposal stage, e.g., YOLOv3 [102] or Single-Shot Multi-boxDetector (SSD) [74], or do not rely on proposals at all [67].

The current approach uses a single intensity image of a driver camera. In extension, the method could employ multiple cameras, such as the in-cabin camera provided by the dataset of Chapter 3, as the trend of holistic driver monitoring shows [57, 85]. A larger observed in-car volume also allows for the analysis of multiple passengers, further broadening the scope to comfort and multi-media applications in SAE level 3 and above. In that case, one can generalize the term *driver observation* to *occupant observation*. There is an abundance of information the human body expresses besides head pose: eye gaze, facial landmarks and their configuration, and body posture, to name a few. Detecting these for can gain understanding about the emotions, vigilance of each individual, their interaction with each other inside the vehicle, with the vehicle they occupy, and with the environment. As soon as occupant observation sensors become a commodity in series vehicles (i.e., at an affordable price as opposed to being part of expensive extras), more advanced functions may be developed in the future for both convenience and entertainment. For instance, occupants can get information about the outside world from the vehicle, like nearby sights or mountain peaks which can be conveniently displayed in an augmented-reality fashion since the location of the occupants eyes are known.

Vehicle Environment Perception

Pedestrians are the main object class of interest in the vehicle's environment within the scope of this thesis. Chapter 5 contributed to the multi-sensor 3D person detection, namely from camera and lidar. It showed improvements by learning raw features instead of hand-crafting, therefore following the trend of end-to-end learning. A component of the architecture that is still manually defined is the anchor proposals that reside on the ground plane. It is therefore dependent on a robust ground plane estimation algorithm. Possible improvements lead in the direction of anchor-free methods, similar to the aforementioned anchor-free extensions of the head pose estimation method.

In Chapter 6 pedestrian head pose was estimated as an additional feature besides the location of the pedestrian. Pedestrian head pose serves as a context cue for obtaining the pedestrian's awareness of the approaching vehicle. One can think of further features of the pedestrian which play useful for understanding its behavior and interaction with the environment. Specifically, 3D body pose (i.e., 3D location of joint points) can be used for obtaining gestures and to serve as a context cue for awareness or the direction of travel. A different direction could be to use additional sensors for environment perception, such as radar. Radar sensors could bring additional context cues by micro-Doppler measurements [137], allowing for instant extraction of pedestrian leg movement. Early fusion of data from multiple sensor modalities which perceive the 360° surrounding of the vehicle may also improve detection performance and robustness to adverse weather. This defines the current trend.

The methods of Chapter 5 and Chapter 6 showcase the perception of pedestrians. Looking beyond the scope of this thesis, any other object which the driver or the pedestrian might interact with is also of interest. E.g., a pedestrian might walk towards a bus station on the other side of the road, or cross based on traffic light status [130]. Similarly, the driver can interact with other traffic participants besides pedestrians, such as cyclists.

DRIVER-ROAD USER INTERACTION FOR PATH PREDICTION

The presented framework for driver-pedestrian path prediction has shown a positive impact of incorporating mutual awareness on the head pose estimates of the road users. The analyzed scenario of a single, potentially crossing pedestrian is, without a doubt, very important. Yet, there are many even more complex interactions in real-life traffic environments. Extensions towards the path prediction of multiple traffic participants are reasonable next steps, especially incorporating their joint awareness of each other. Dependencies amongst pairs of road users could be added to the DBN, but limiting them to close spatial proximity, to remain scalable with increasing number of road users. The DBN architecture of Chapter 6 has computational limitations in scaling up the number of traffic participants, mainly restricted by the number of edges between discrete states. After all, it might be beneficial to learn the dependencies from data, which are hand-crafted for now. To that end Girase *et al.* [44] propose to use a data-driven multi-agent approach with long-term goal prediction, short-term intention classification and optimization of paths over a scene graph. This would counteract one limitation of the DBN-based "open loop" system which predicts future paths based on observations from the past, and does not consider future actions and interactions the traffic participants might undergo. A further potential direction could be to model agent interactions in a graph structure within a neural network [3].

The work of Chapter 6 showed a synergy between expert knowledge and a data-driven approach. The former enables explainability (i.e., allow for introspecting the intermediate representations, such as the latent awareness states), while the latter uses data to optimize the parameters. There is expert knowledge integrated into the DBN, such as switching linear vehicle motion models and switching linear pedestrian motion models. Similarly, mutual awareness is a crafted attribute of the model and could be further improved by more sophisticated models of driver and pedestrian awareness (e.g., fixation cues), potentially extracted from data and over time, e.g., by attention networks [24, 130]. A clear goal would be "exchanged" awareness [132], i.e., modeling the driver's belief about the pedestrian's awareness in addition to the driver's awareness of the pedestrian's presence.

There is a balance to find, as purely data-driven models based on convolutional neural networks (CNNs) [103], LSTMs [3] or GANs [48] might find suitable features (spatially and temporally), yet at the lack of explainability. Going forward, end-to-end models trained along the full functional chain from raw sensor data up to predicting future behavior distributions of multiple agents [142] define the current trend. An ideal multi-agent path prediction system would allow to encode expert knowledge where needed, restrict a subset of intermediate representations to be interpretable (e.g. motion states), and allows all other parts to be learned from representative data.

AN END-TO-END SYSTEM

Some common themes transcend the individual chapters of this thesis. For instance, end-to-end training has been employed in Chapter 4 and Chapter 5, and features and internal representations have been learned in the differentiable models of Chapter 4, Chapter 5 and Chapter 6. Note that the DBN of Chapter 6 is differentiable despite not being a neural network. These themes can be extended along the whole processing chain of driver observation, environment perception and path prediction. Therefore,

the following two aspects will be beneficial for overall system performance in future systems. The first key aspect is developing end-to-end optimization for a system that predicts behaviors and probabilistic paths of traffic participants (including the driver) from multi-modal sensor input (including the vehicle's interior). This requires a full differentiable model which optimizes all intermediate representations directly towards the end objective. Intermediate representation optimized in isolation might be optimal for an intermediate objective (e.g., pedestrian locations), but not necessarily for the quality of the overall system predictions.

The second key aspect is uncertainty: the DBN performs probabilistic path prediction. Extending the input components, such as the driver head pose estimation, as well as the pedestrian measurements to also estimate uncertainties will improve the overall handling over uncertainty up to the predicted paths (i.e., the predicted paths will be more spatially spread in case of larger uncertainties). Overall, these attributes will likely show benefits in the temporal aspect of the predictions, as time-series can be used during optimization of the end-to-end system. Other important aspects for realizing such an end-to-end system are data sufficiency and data quality. They are discussed next.

Data to Accelerate Automated Driving

While this thesis made contributions towards better driver head pose, person detection, and driver-pedestrian path prediction for ADAS in intelligent vehicles, there is still a long way to go for the complex task of automated driving in urban areas. Data is a key factor in developing and evaluating intelligent vehicles.

The driver head pose dataset *DD-Pose* presented in this thesis provides a large number of head pose annotations and allowed for robust head pose estimation performance. Despite being large in size (330 k images), still, only a small number of different persons are present in the dataset. The experiments have shown a good generalization of the model to the unseen persons of the test subset. Yet, showing generalization on a broader spectrum of age and ethnicity would be favorable. Towards that end, there is the challenge of obtaining ground truth head pose annotations, which need to be acquired invasively, i.e., by a head-worn device. This thesis presented a quick-to-calibrate head pose measurement device. Increasing the usability by reducing the per-subject calibration effort further could simplify fleet data collection that is needed to obtain head-pose data from potentially hundred or thousands of different drivers. Since the release of *DD-Pose*, other driver-related datasets have been published. They follow the trend towards multiple cameras capturing the driver in a more holistic view, also covering body and hands from multiple perspectives [57, 85] and also recording audio inside the vehicle [57]. Some follow a hybrid approach of naturalistic driving and simulator data [85]. Besides estimating head-pose, the focus goes towards estimating driver attention, alertness and behavior recognition.

Only a few publicly available driver-based datasets have integrated sensor data capturing the surrounding of the vehicle (including the dataset of Chapter 3). A challenge arises when data is incomplete in terms of sensor modality (e.g., missing lidar or radar data), temporal (e.g., sparse sampling of scenes), perceived environment (e.g., missing driver observation), or traffic scenarios (e.g., underrepresented pedestrian interactions), to name a few. Lacking a 'universal' dataset, the work of this thesis worked with three disjoint datasets: the KITTI dataset [43], containing exterior camera and lidar data was

used for developing the 3D person detector (Chapter 5), lacking in-cabin sensing. The *DD-Pose* dataset (Chapter 3) was created in the course of this thesis to develop a head pose estimation method (Chapter 4). While it provides images from a camera pointing outside, its scenarios covered naturalistic driving in urban scenes, yet with a small number of pedestrian crossing interactions. To that end, yet another dataset was collected to advance driver-pedestrian path prediction (Chapter 6), containing data from inside the vehicle (driver camera) and outside the vehicle, and focusing on pedestrian-crossing scenarios. Obtaining the data required effortful buildup of a research vehicle, data selection, preprocessing, and annotation. Going forward, the scientific community would greatly benefit from representative, worldwide, naturalistic, multi-sensor, temporal data which cover the outside environment as well as the inside of the vehicle. Great advances in that direction have already been made in the research community by publicly releasing ECP2.5D [11], View-of-Delft [88], NuScenes [15], ArgoVerse2[136], and the Waymo Open Dataset [34], despite lacking interior sensing.

When it comes to datasets for neural networks in general, it is commonly believed that 'more is better'. While it has been shown that network performance increases with the amount of training data, it also comes with the cost of potentially increased training time. To some degree this can be mitigated by having a network with good generalization capabilities, which allows reduction of training epochs with increased training data (i.e., how often will a single training sample be used during the learning process). So from a perspective of "maximum tolerated training time", the problem becomes: "which data has the most value in training to yield a good model?". To that end, smart data collection and sub-selection strategies are needed, such as active learning approaches [104] which support selecting data with rare and difficult cases. Perhaps rare and difficult cases can best be addressed by collecting such data through generative models or advanced simulations. This becomes especially important for collision scenarios, as was assimilated in the dataset of Chapter 6.

The experience gained through the creation of this thesis leads to the following three proposals for the community working towards automated driving, may it be researchers, suppliers, or OEMs with the goal of earlier releases of better systems. First, data sharing: the large predicted market of autonomous driving has motivated suppliers and OEMs to invest in worldwide data collection, yet keeping data private, partly for competitive reasons. Imagining vehicles from multiple competitors recording the same geolocation and investing in annotating the same scenes motivates the thought of sharing data to make a larger earlier societal impact. An additional factor to consider when recording data in public and sharing among parties is handling data protection and privacy regulations, such as general data protection regulation (GDPR). Second, data standardization: data sharing demands a unified standard of data representation (raw sensor data, but also annotations), as Robot Operating System (ROS) has achieved in part [42]. Third, central storage: a central storage to share data with a unified API to access and add new interesting data could further speed up development, and make it accessible for research institutions that have not committed to investing in building up an own research vehicle.

Yet, the automotive industry might need a legislative framework to follow this track.

Summarizing, this thesis presented a framework looking-in and looking-out an intel-

ligent vehicle to perceive driver and pedestrians. It led to advancements in head pose estimation, 3D person detection, and path prediction. It combined the perception of the outside and the inside of the vehicle. Future challenges have been identified alongside ideas improving the proposed methods to further contribute to understanding the behavior of humans inside and outside an intelligent vehicle.

It is the author's hope that continuation of the research in this thesis will lead to safer ADAS for upcoming consumer vehicles, as well as fully self-driving vehicles of the future.

ACKNOWLEDGEMENTS

I am very thankful for the support, guidance, and encouragement of many special people that have been by my side during the creation of this thesis. First and foremost, I would like to sincerely thank my advisor and promotor Prof. Dr. Dariu M. Gavrila. Thank you Dariu, for your excellent advice and guidance, and for the critical questions that lead to a high quality of work. You helped me to push my own limits - both in science and on ski slopes.

Additionally, I would like to convey my gratitude to my co-promotor Dr. Julian F. P. Kooij, as well as the committee members, for their valuable feedback and diligent review of this thesis.

Many of my colleagues at Mercedes-Benz and the TU Delft were instrumental in the development of this thesis. I am deeply thankful to Dr. Ulrich Kressel for his appreciation of my work, for providing opportunities and guard rails, and for supporting my ideas. Special thanks to Prof. Dr. Fabian Flohr for the collaboration on one of my first publications. I appreciate your creative vision and I am grateful for your encouragement. Dr. Jork Stapel, thank you for our close collaboration, in Delft, in Stuttgart, and online. Your energy, pragmatism, and creative drive are a great combination to build more great things. One of my deepest gratitude goes toward my colleagues and close friends Dr. Markus Braun and Sebastian Krebs. You are some of the smartest and most humorous people I know. The productive hacking sessions with you are always an amazing experience. Even more, I have enjoyed our activities outside of work very much: in the mountains both during summer time and winter time, but also on the Soča and the Danube river. I had the great pleasure to collaborate with several students, in particular Stephan Siebold, Dominik Jargot, and Julia Reuss. Building the driver observation hardware provided a great balance to programming. I can still smell the soldering fumes when remembering our late-evening tinkering. A special thank you goes to Dr. Andreas Fregin, for evolving from colleagues to the closest friends. Your fast footsteps on the corridor showed me early on that you're driven and ambitious. I am very thankful for introducing me to the Robot Operating System (ROS) framework, which enabled many parts of this thesis, particularly the dynamic coordinate transforms involved in head pose estimation. Besides, I enjoy that we share our hobbies of tinkering and maintaining cars. When we are working together, we always seem to be a step ahead of each other, which makes the satisfaction of having drinks after completing a job even more enjoyable.

I am immensely grateful to my friends and family, who have supported me in my personal life and without whom I would not have been able to accomplish this endeavor. In particular, my close friends Markus Helwig, Micha Bruns, Philipp Tölle, and Julia Reinsch who have always supported me, also by challenging my quirks. I also want to express my deepest gratitude to my family, Reinhold, Sunhild, and Sonja. Reinhold, you taught me to create things and to be precise when doing so. I am certain that our time fixing the house in Romania, but also tinkering with electronics build the foundation of

my career. Your expertise is invaluable. Thank you Sunhild, all my life you have given me unconditional emotional support, shared your wisdom, and reminded me of the important things in life. Sonja, thank you also for your support, you are my role model for success and contentment. Finally, Stephanie – my warmest appreciation goes to you. Thank you for being by my side, for your understanding, patience, and encouragement. You provided me with the best balance I could imagine. Most importantly, you taught me that happiness not only lies in creating things but also in enjoying the beauty that nature created for us.

CURRICULUM VITÆ

Markus ROTH

1986-07-19 Born in Talmesch, Romania.

EDUCATION

2014–present Ph.D. in the Cognitive Robotics department
Delft University of Technology, The Netherlands
Thesis: Driver and Pedestrian Mutual Awareness for Path Prediction in Intelligent Vehicles

2007–2014 Dipl.-Inform., Computer Science
Karlsruhe Institute of Technology, Germany
graduated with distinction

2011–2012 Research semester at USC IRIS Computer Vision Lab
Study Thesis grant-aided by the interACT Scholarship
Supervisors: Ram Nevatia (USC), and Rainer Stiefelhagen (KIT)
University of Southern California, Los Angeles, USA

PROFESSIONAL EXPERIENCE

2014–present Mercedes-Benz AG, Germany
Machine Learning Engineer in the department Perception and Maps
Environment perception of Vulnerable Road Users

2012–2014 Videmo Intelligent Video Analysis GmbH, Karlsruhe, Germany
Software Engineer for video analysis

List of Publications

Publications

7. **M. Roth** and D. M. Gavrila, *"intrApose: Monocular Driver 6 DOF Head Pose Estimation Leveraging Camera Intrinsics"*, IEEE Trans. on Intelligent Vehicles (TIV), 2023, vol. 8, no. 8, pp. 4057–4068.

   Author contributions: M. Roth designed, implemented and evaluated the proposed method. M. Roth wrote the journal article. D. M. Gavrila provided guidance and supervision.

6. M. Upreti, J. Ramesh, C. Kumar, B. Chakraborty, V. Balisavira, **M. Roth**, V. Kaiser and P. Czech, *"Uncertainty and Traffic Light Aware Pedestrian Crossing Intention Prediction"*, IEEE Intelligent Vehicles Symposium Workshops (IV Workshops), 2023.

   Author contributions: M. Upreti, J. Ramesh, C. Kumar, and B. Chakraborty designed, implemented and evaluated the proposed method, and wrote the paper. V. Balisavira, M. Roth, V. Kaiser and P. Czech provided guidance and supervision.

5. **M. Roth**, J. Stapel, R. Happee, and D. M. Gavrila, *"Driver and Pedestrian Mutual Awareness for Path Prediction and Collision Risk Estimation"*, IEEE Trans. on Intelligent Vehicles (TIV), 2022, vol. 7, no. 4, pp. 896–907.

   Author contributions: M. Roth contributed substantially to the concept, training, and evaluation of the method. M. Roth designed the model architecture, vehicle instrumentation, and evaluation methodology and refined it collaboratively with J. Stapel. J. Stapel built up the sensor vehicle. M. Roth and J. Stapel recorded the data and conducted the experiments. M. Roth and J. Stapel wrote the journal article in close collaboration. R. Happee contributed to the design of the study and review of the manuscript. D. M. Gavrila contributed to the concept and writing of the journal article and provided guidance and supervision.

4. **M. Roth**, D. Jargot, and D. M. Gavrila, *Deep End-to-end 3D Person Detection from Camera and Lidar*, Proc. of the IEEE Intelligent Transportation Systems Conference, 2019, pp. 521–527.

   Author contributions: M. Roth conducted the experiments, evaluated the method and wrote the paper, together with D. Jargot. D. Jargot implemented the proposed method. D. M. Gavrila provided guidance and supervision.

3. **M. Roth** and D. M. Gavrila, *DD-Pose - A large-scale Driver Head Pose Benchmark*, Proc. of the IEEE Intelligent Vehicles Symposium, 2019, pp. 927–934.

   Author contributions: M. Roth developed the sensor setup, calibration methodology, built up the research vehicle, acquired and processed the data, conducted the experiments, and wrote the paper. M. Roth and D. M. Gavrila designed the driving scenarios. D. M. Gavrila furthermore provided guidance and supervision.

2. A. Fregin, **M. Roth**, M. Braun, S. Krebs, and F. Flohr, *Building a Computer Vision Research Vehicle with ROS*, Proc. of the ROSCon, 2017.

   Author contributions: All authors have contributed to building several computer vision research vehicles with ROS. A. Fregin created the publication slides.

1. **M. Roth**, F. Flohr, and D. M. Gavrila, *Driver and Pedestrian awareness-based Collision Risk Analysis*, Proc. of the IEEE Intelligent Vehicles Symposium, 2016, pp. 454–459.

   Author contributions: M. Roth created, implemented and evaluated the proposed method. F. Flohr worked on the visual feature extraction of the pedestrian and helped with data acquisition and the technical and mathematical formulation of the model. M. Roth, F. Flohr, and D. M. Gavrila wrote the paper together. D. M. Gavrila furthermore provided guidance and supervision.

PATENTS

2. **M. Roth**, R. Ivancevic, W. Stolzmann, *Verfahren sowie System zum Ermitteln von Raumkoordinaten von Landmarken eines Kopfes einer Person*,
   German Patent No. DE102018002224A1, 2018.

1. **M. Roth**, *Verfahren zur Bestimmung einer Kopfpose eines Kopfes*,
   German Patent No. DE102017000962A1, 2017.

BIBLIOGRAPHY

[1] A. F. Abate, C. Bisogni, A. Castiglione, and M. Nappi. “Head pose estimation: An extensive survey on recent techniques and applications”. In: *Pattern Recognition* 127 (2022), p. 108591.

[2] B. Ahn, D. G. Choi, J. Park, and I. S. Kweon. “Real-time head pose estimation using multi-task deep neural network”. In: *Robotics and Autonomous Systems* 103 (2018), pp. 1–12.

[3] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. “Social LSTM: Human Trajectory Prediction in Crowded Spaces”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 961–971.

[4] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner. “img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 7617–7627.

[5] M. Ariz, J. J. Bengoechea, A. Villanueva, and R. Cabeza. “A novel 2D/3D database with automatic face annotation for head tracking and pose estimation”. In: *Computer Vision and Image Understanding (CVIU)* 148 (2016), pp. 201–210.

[6] T. Baltrusaitis, P. Robinson, and L. P. Morency. “3D Constrained Local Model for rigid and non-rigid facial tracking”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 2610–2617.

[7] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. “OpenFace 2.0: Facial Behavior Analysis Toolkit”. In: *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)*. 2018, pp. 59–66.

[8] J. M. D. Barros, B. Mirbach, F. Garcia, K. Varanasi, and D. Stricker. “Fusion of Keypoint Tracking and Facial Landmark Detection for Real-Time Head Pose Estimation”. In: *IEEE Winter Conf. on Applications of Computer Vision (WACV)*. 2018, pp. 2028–2037.

[9] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara. “Face-from-Depth for Head Pose Estimation on Depth Images”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 42.3 (2020), pp. 596–609.

[10] C. Braeuchle, J. Ruenz, F. Flehmig, W. Rosenstiel, and T. Kropf. “Situation analysis and decision making for active pedestrian protection using Bayesian networks”. In: *6. Tagung Fahrerassistenzsysteme, TÜV SÜD*. 2013, pp. 1–5.

[11] M. Braun, S. Krebs, and D. M. Gavrila. “ECP2.5D - Person Localization in Traffic Scenes”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2020, pp. 1694–1701.

[12] M. Braun, S. Krebs, F. Flohr, and D. Gavrila. “EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 41.8 (2019), pp. 1844–1861.

[13] M. Braun, Q. Rao, Y. Wang, and F. Flohr. “Pose-RCNN: Joint Object Detection and Pose Estimation Using 3D Object Proposals”. In: *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*. 2016, pp. 1546–1551.

[14] N. Brouwer, H. Kloeden, and C. Stiller. “Comparison and Evaluation of Pedestrian Motion Models for Vehicle Safety Systems”. In: *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*. 2016, pp. 2207–2212.

[15] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. “nuScenes: A multimodal dataset for autonomous driving”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11621–11631.

[16] Z. Cao, Z. Chu, D. Liu, and Y. Chen. “A Vector-based Representation to Enhance Head Pose Estimation”. In: *IEEE Winter Conf. on Applications of Computer Vision (WACV)*. 2021, pp. 1188–1197.

[17] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. Chateau. “Deep MANTA: A Coarse-to-Fine Many-Task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1827–1836.

[18] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. “Deep, Landmark-Free FAME: Face Alignment, Modeling, and Expression Estimation”. In: *Int. Journal of Computer Vision (IJCV)* 127.6 (2019), pp. 930–956.

[19] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 77–85.

[20] I. Chatterjee, Isha, and A. Sharma. “Driving Fitness Detection: A Holistic Approach For Prevention of Drowsy and Drunk Driving using Computer Vision Techniques”. In: *South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference*. 2018, pp. 1–6.

[21] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. “Monocular 3D Object Detection for Autonomous Driving”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2147–2156.

[22] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. “3D Object Proposals for Accurate Object Class Detection”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2015, pp. 424–432.

[23] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. “Multi-view 3D Object Detection Network for Autonomous Driving”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6526–6534.

[24] X. Chen, Z. Wu, and J. Yu. “TSSD: Temporal Single-Shot Detector Based on Attention and LSTM”. In: *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. 2018, pp. 1–9.

[25] J. B. Cicchino. “Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates”. In: *Accident Analysis & Prevention* 99 (2017), pp. 142–152.

[26] J. Dahl, G. R. de Campos, C. Olsson, and J. Fredriksson. “Collision Avoidance: A Literature Review on Threat-Assessment Techniques”. In: *IEEE Trans. on Intelligent Vehicles (TIV)* 4.1 (2019), pp. 101–113.

[27] J. Dai, L. Yi, K. He, and J. Sun. “R-FCN: Object Detection via Region-based Fully Convolutional Networks”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2016, pp. 379–387.

[28] G. De Nicolao, A. Ferrara, and L. Giacomini. “Onboard Sensor-Based Collision Risk Assessment to Improve Pedestrians’ Safety”. In: *IEEE Trans.. on Vehicular Technology* 56.5 (2007), pp. 2405–2413.

[29] D. F. DeMenthon and L. S. Davis. “Model-based object pose in 25 lines of code”. In: *European Conf. on Computer Vision (ECCV)*. 1992, pp. 335–343.

[30] N. Deo and M. M. Trivedi. “Looking at the Driver/Rider in Autonomous Vehicles to Predict Take-Over Readiness”. In: *IEEE Trans. on Intelligent Vehicles (TIV)* 5.1 (2020), pp. 41–52.

[31] D. Derkach, A. Ruiz, and F. M. Sukno. “Tensor Decomposition and Non-linear Manifold Modeling for 3D Head Pose Estimation”. In: *Int. Journal of Computer Vision (IJCV)* 127.10 (2019), pp. 1565–1585.

[32] K. Diaz-Chito, A. Hernández-Sabaté, and A. M. López. “A reduced feature set for driver head pose estimation”. In: *Applied Soft Computing* 45 (2016), pp. 98–107.

[33] A. El Khatib, C. Ou, and F. Karray. “Driver Inattention Detection in the Context of Next-Generation Autonomous Vehicles Design: A Survey”. In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* 21.11 (2020), pp. 4483–4496.

[34] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov. “Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset”. In: *IEEE Int. Conf. on Computer Vision (ICCV)*. 2021, pp. 9710–9719.

[35] Euro NCAP. “Test Protocol – AEB/LSS VRU systems v4.4”. In: *European New Car Assessment Programme (Euro NCAP)* (2023), pp. 1–72.

[36] European Commission. “Road safety thematic report – Pedestrians”. In: *European Commission, Directorate General for Transport* (2021), p. 18.

[37] European Commission. “Traffic Safety Basic Facts on Junctions”. In: *European Commission, Directorate General for Transport* (2018), p. 21.

[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. “The Pascal Visual Object Classes Challenge”. In: *Int. Journal of Computer Vision (IJCV)* 88.2 (2010), pp. 303–338.

[39] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. “Random Forests for Real Time 3D Face Analysis”. In: *Int. Journal of Computer Vision (IJCV)* 101.3 (2013), pp. 437–458.

[40] L. Ferranti, B. Brito, E. Pool, Y. Zheng, R. M. Ensing, R. Happee, B. Shyrokau, J. F. P. Kooij, J. Alonso-Mora, and D. M. Gavrila. “SafeVRU: A Research Platform for the Interaction of Self-Driving Vehicles with Vulnerable Road Users”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2019, pp. 1660–1666.

[41] A. Firintepe, M. Selim, A. Pagani, and D. Stricker. “The More, the Merrier? A Study on In-Car IR-based Head Pose Estimation”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2020, pp. 1060–1065.

[42] A. Fregin, M. Roth, M. Braun, S. Krebs, and F. Flohr. “Building a Computer Vision Research Vehicle with ROS”. In: *Proc. of the ROSCon*. 2017.

[43] A. Geiger, P. Lenz, and R. Urtasun. “Are we ready for autonomous driving? The KITTI vision benchmark suite”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 3354–3361.

[44] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi. “Loki: Long term and key intentions for trajectory prediction”. In: *IEEE Int. Conf. on Computer Vision (ICCV)*. 2021, pp. 9803–9812.

[45] R. Girshick. “Fast R-CNN”. In: *IEEE Int. Conf. on Computer Vision (ICCV)*. 2015, pp. 1440–1448.

[46] R. Girshick, J. Donahue, T. Darrell, and J. Malik. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 580–587.

[47] J. Gu, X. Yang, S. De Mello, and J. Kautz. “Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1531–1540.

[48] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. “Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 2255–2264.

[49] S. Gupta, M. Vasardani, and S. Winter. “Negotiation between Vehicles and Pedestrians for the Right of Way at Intersections”. In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* 20.3 (2019), pp. 888–899.

[50] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[51] D. Helbing and P. Molnár. “Social force model for pedestrian dynamics”. In: *Physical Review E* 51.5 (1995), pp. 4282–4286.

[52] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi. “6D Rotation Representation For Unconstrained Head Pose Estimation”. In: *IEEE Int. Conf. on Image Processing (ICIP)*. 2022, pp. 2496–2500.

[53] H. Hirschmüller. “Stereo Processing by Semi-Global Matching and Mutual Information”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 30.2 (2008), pp. 328–341.

[54] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee. "QuatNet: Quaternion-Based Head Pose Estimation With Multiregression Loss". In: *IEEE Trans. on Multimedia* 21.4 (2019), pp. 1035–1046.

[55] B. Huang, R. Chen, W. Xu, and Q. Zhou. "Improving head pose estimation using two-stage ensembles with top-k regression". In: *Image and Vision Computing* 93 (2020), p. 103827.

[56] D. Jargot. "Deep End-to-end Network for 3D Object Detection in the Context of Autonomous Driving". In: *MS Thesis TU Delft* (2019).

[57] S. Jha, M. F. Marzban, T. Hu, M. H. Mahmoud, N. Al-Dhahir, and C. Busso. "The Multimodal Driver Monitoring Database: A Naturalistic Corpus to Study Driver Attention". In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* 23.8 (2022), pp. 10736–10752.

[58] C. G. Keller and D. M. Gavrila. "Will the Pedestrian Cross? A Study on Pedestrian Path Prediction". In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* 15.2 (2014), pp. 494–506.

[59] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *Int. Conference on Learning Representations (ICLR)*. 2015, pp. 1–15.

[60] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. "Context-based Pedestrian Path Prediction". In: *European Conf. on Computer Vision (ECCV)*. 2014, pp. 618–633.

[61] J. F. P. Kooij, F. Flohr, E. A. I. Pool, and D. M. Gavrila. "Context-Based Path Prediction for Targets with Switching Dynamics". In: *Int. Journal of Computer Vision (IJCV)* 127.3 (2019), pp. 239–262.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems (NIPS)* 25 (2012), pp. 1–9.

[63] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander. "Joint 3D Proposal Generation and Object Detection from View Aggregation". In: *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. 2018, pp. 1–8.

[64] M. La Cascia, S. Sclaroff, and V. Athitsos. "Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 22.4 (2000), pp. 322–336.

[65] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom. "PointPillars: Fast Encoders for Object Detection from Point Clouds". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 12697–12705.

[66] S. Lathuiliere, P. Mesejo, X. Alameda-Pineda, and R. Horaud. "A Comprehensive Analysis of Deep Regression". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 42.9 (2020), pp. 2065–2081.

[67] H. Law and J. Deng. "CornerNet: Detecting objects as paired keypoints". In: *European Conf. on Computer Vision (ECCV)*. 2018, pp. 734–750.

[68] S. Lefèvre, D. Vasquez, and C. Laugier. “A survey on motion prediction and risk assessment for intelligent vehicles”. In: *ROBOMECH Journal* 1.1 (2014), pp. 1–14.

[69] J. Levinson, C. Esteves, K. Chen, N. Snavely, A. Kanazawa, A. Rostamizadeh, and A. Makadia. “An Analysis of SVD for Deep Rotation Estimation”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020, pp. 22554–22565.

[70] B. Li, T. Zhang, and T. Xia. “Vehicle Detection from 3D Lidar Using Fully Convolutional Network”. In: *Robotics: Science and Systems XII*. 2016, pp. 1–8.

[71] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao, et al. “LoGoNet: Towards Accurate 3D Object Detection with Local-to-Global Cross-Modal Fusion”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 17524–17534.

[72] Y. Li, X.-Y. Lu, J. Wang, and K. Li. “Pedestrian Trajectory Prediction Combining Probabilistic Reasoning and Sequence Learning”. In: *IEEE Trans. on Intelligent Vehicles (TIV)* 5.3 (2020), pp. 461–474.

[73] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. “Feature Pyramid Networks for Object Detection”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 936–944.

[74] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. “SSD: Single shot multibox detector”. In: *European Conf. on Computer Vision (ECCV)*. 2016, pp. 21–37.

[75] S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, and M. Trivedi. “On the design and evaluation of robust head pose for visual user interfaces”. In: *Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications*. 2012, pp. 149–154.

[76] S. Martin, K. Yuen, and M. M. Trivedi. “Vision for Intelligent Vehicles & Applications (VIVA): Face detection and head pose challenge”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2016, pp. 1010–1014.

[77] T. P. Minka. “A family of algorithms for approximate Bayesian inference”. PhD thesis. Massachusetts Institute of Technology, 2001.

[78] P. Molchanov, S. Gupta, K. Kim, and J. Kautz. “Hand gesture recognition with 3D convolutional neural networks”. In: *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2015, pp. 1–7.

[79] T. Moore and D. Stouch. “A Generalized Extended Kalman Filter Implementation for the Robot Operating System”. In: *Intelligent Autonomous Systems 13*. 2016, pp. 335–348.

[80] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. “3D Bounding Box Estimation Using Deep Learning and Geometry”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5632–5640.

[81] E. Murphy-Chutorian and M. M. Trivedi. “Head Pose Estimation in Computer Vision: A Survey”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 31.4 (2009), pp. 607–626.

[82] M. I. Nazir, K. M. A. Al Razi, Q. S. Hossain, and S. K. Adhikary. “Pedestrian Flow Characteristics At Walkways In Rajshahi Metropolitan City Of Bangladesh”. In: *Int. Conf. on Civil Engineering for Sustainable Development*. 2014, pp. 978–984.

[83] S. Neogi, M. Hoy, K. Dang, H. Yu, and J. Dauwels. “Context Model for Pedestrian Intention Prediction Using Factored Latent-Dynamic Conditional Random Fields”. In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* 22.11 (2021), pp. 6821–6832.

[84] J. Nuevo, L. M. Bergasa, and P. Jiménez. “RSMAT: Robust simultaneous modeling and tracking”. In: *Pattern Recognition Letters* 31.16 (2010), pp. 2455–2463.

[85] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado. “DMD: A Large-Scale Multi-modal Driver Monitoring Dataset for Attention and Alertness Analysis”. In: *European Conf. on Computer Vision Workshops (ECCVW)*. 2020, pp. 387–405.

[86] A. Palffy, J. Dong, J. F. P. Kooij, and D. M. Gavrila. “CNN Based Road User Detection using the 3D Radar Cube”. In: *IEEE Robotics and Automation Letters (RA-L)* 5.2 (2020), pp. 1263–1270.

[87] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila. “Multi-Class Road User Detection With 3+1D Radar in the View-of-Delft Dataset”. In: *IEEE Robotics and Automation Letters (RA-L)* 7.2 (2022), pp. 4961–4968.

[88] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila. “Multi-class road user detection with 3+1D radar in the View-of-Delft dataset”. In: *IEEE Robotics and Automation Letters (RA-L)* 7.2 (2022), pp. 4961–4968.

[89] J. Paone, D. Bolme, R. Ferrell, D. Aykac, and T. Karnowski. “Baseline face detection, head pose estimation, and coarse direction detection for facial data in the SHRP2 naturalistic driving study”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2015, pp. 174–179.

[90] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. “You’ll never walk alone: Modeling social behavior for multi-target tracking”. In: *IEEE Int. Conf. on Computer Vision (ICCV)*. 2009, pp. 261–268.

[91] V. Peretroukhin, M. Giamou, W. N. Greene, D. Rosen, J. Kelly, and N. Roy. “A Smooth Representation of Belief over SO(3) for Deep Rotation Learning with Uncertainty”. In: *Robotics: Science and Systems XVI*. 2020, pp. 1–9.

[92] T. Pham, M. Maghoumi, W. Jiang, B. S. S. Jujjavarapu, M. S. X. Liu, H.-C. Lin, B.-J. Chen, G. Truong, C. Fang, J. Kwon, and M. Park. *NVAutoNet: Fast and Accurate 360° 3D Visual Perception For Self Driving*. 2023.

[93] E. A. I. Pool, J. F. P. Kooij, and D. M. Gavrila. “Crafted vs. Learned Representations in Predictive Models - A Case Study on Cyclist Path Prediction”. In: *IEEE Trans. on Intelligent Vehicles (TIV)* 6.4 (2021), pp. 747–759.

[94] E. A. I. Pool, J. F. P. Kooij, and D. M. Gavrila. “Using Road Topology to Improve Cyclist Path Prediction”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2017, pp. 289–296.

[95] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. “Frustum PointNets for 3D Object Detection from RGB-D Data”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 918–927.

[96] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 30. 2017, pp. 1–10.

[97] R. Qian, X. Lai, and X. Li. “3D Object Detection for Autonomous Driving: A Survey”. In: *Pattern Recognition* 130.1 (2022), p. 108796.

[98] R. Quintero, I. Parra, D. F. Llorca, and M. A. Sotelo. “Pedestrian Intention and Pose Prediction through Dynamical Models and Behaviour Classification”. In: *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*. 2015, pp. 83–88.

[99] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. “An All-In-One Convolutional Neural Network for Face Analysis”. In: *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)*. 2017, pp. 17–24.

[100] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788.

[101] J. Redmon and A. Farhadi. “YOLO9000: Better, Faster, Stronger”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6517–6525.

[102] J. Redmon and A. Farhadi. “YOLOv3: An Incremental Improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).

[103] E. Rehder, F. Wirth, M. Lauer, and C. Stiller. “Pedestrian Prediction by Planning Using Deep Neural Networks”. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2018, pp. 5903–5908.

[104] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, X. Chen, and X. Wang. “A survey of deep active learning”. In: *ACM Computing Surveys* 54.9 (2021), pp. 1–40.

[105] S. Ren, K. He, R. Girshick, and J. Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 39.6 (2017), pp. 1137–1149.

[106] D. Ridel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf. “A Literature Review on the Prediction of Pedestrian Behavior in Urban Scenarios”. In: *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*. 2018, pp. 3105–3112.

[107] M. Roth, F. Flohr, and D. M. Gavrila. “Driver and Pedestrian Awareness-based Collision Risk Analysis”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2016, pp. 454–459.

[108] M. Roth and D. M. Gavrila. “intrApose: Monocular Driver 6 DOF Head Pose Estimation Leveraging Camera Intrinsics”. In: *IEEE Trans. on Intelligent Vehicles (TIV)* 8.8 (2023), pp. 4057–4068.

[109] M. Roth and D. M. Gavrila. “DD-Pose - A large-scale Driver Head Pose Benchmark”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2019, pp. 927–934.

[110] M. Roth, D. Jargot, and D. M. Gavrila. “Deep End-to-end 3D Person Detection from Camera and Lidar”. In: *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*. 2019, pp. 521–527.

[111] M. Roth, J. Stapel, R. Happee, and D. M. Gavrila. “Driver and Pedestrian Mutual Awareness for Path Prediction and Collision Risk Estimation”. In: *IEEE Trans. on Intelligent Vehicles (TIV)* 7.4 (2022), pp. 896–907.

[112] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. “Human motion trajectory prediction: a survey”. In: *Int. Journal of Robotics Research (IJRR)* 39.8 (2020), pp. 895–935.

[113] N. Ruiz, E. Chong, and J. M. Rehg. “Fine-Grained Head Pose Estimation Without Keypoints”. In: *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, pp. 2187–2196.

[114] SAE On-Road Automated Driving (ORAD) Committee. “Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles”. In: *SAE International* (2021), pp. 1–41.

[115] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. “300 faces in-the-wild challenge: The first facial landmark Localization Challenge”. In: *IEEE Int. Conf. on Computer Vision Workshops (ICCV Workshops)*. 2013, pp. 397–403.

[116] H. Saptoadi. “Suitable Deceleration Rates for Environmental Friendly City Driving”. In: *Int. Journal of Research in Chemical, Metallurgical and Civil Engineering* 4.1 (2017), pp. 2–5.

[117] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. “Bosphorus Database for 3D Face Analysis”. In: *Biometrics and Identity Management*. 2008, pp. 47–56.

[118] N. Schneider and D. M. Gavrila. “Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study”. In: *German Conf. on Pattern Recognition (DAGM GCPR)*. 2013, pp. 174–183.

[119] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen. “DriveAHead - A Large-Scale Driver Head Pose Dataset”. In: *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017, pp. 1165–1174.

[120] A. Schwarz, Z. Lin, and R. Stiefelhagen. “HeHOP: Highly efficient head orientation and position estimation”. In: *IEEE Winter Conf. on Applications of Computer Vision (WACV)*. 2016, pp. 1–8.

[121] L. Sheng, J. Cai, T.-J. Cham, V. Pavlovic, and K. N. Ngan. “Visibility Constrained Generative Model for Depth-based 3D Facial Pose Tracking”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 41.8 (2019), pp. 1994–2007.

[122] M. D. Shuster. “A Survey of Attitude Representations”. In: *Journal of the Astronautical Sciences* 41.4 (1993), pp. 439–517.

[123] G. Sikander and S. Anwar. “Driver Fatigue Detection Systems: A Review”. In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* 20.6 (2019), pp. 2339–2352.

[124] M. Simon, S. Milz, K. Amende, and H.-M. Gross. “Complex-YOLO: Real-time 3D Object Detection on Point Clouds”. In: *arXiv preprint arXiv:1803.06199* (2018).

[125] K. Simonyan and A. Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *Int. Conference on Learning Representations (ICLR)*. 2015, pp. 1–14.

[126] J. R. van der Sluis, E. A. I. Pool, and D. M. Gavrila. “An Experimental Study on 3D Person Localization in Traffic Scenes”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2020, pp. 1813–1818.

[127] S. Söntges, M. Koschi, and M. Althoff. “Worst-case Analysis of the Time-To-React Using Reachable Sets”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2018, pp. 1891–1897.

[128] J. Stapel, M. E. Hassnaoui, and R. Happee. “Measuring Driver Perception: Combining Eye-Tracking and Automated Road Scene Perception”. In: *Human Factors* 64.4 (2020), pp. 714–731.

[129] L. Tran and X. Liu. “On Learning 3D Face Morphable Model from In-the-wild Images”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 43.1 (2021), pp. 157–171.

[130] M. Upreti, J. Ramesh, C. R. Kumar, B. Chakraborty, V. Balisavira, V. Kaiser, M. Roth, and P. Czech. “Uncertainty and Traffic Light Aware Pedestrian Crossing Intention Prediction”. In: *IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*. 2023.

[131] M. Venturelli, G. Borghi, R. Vezzani, and R. Cucchiara. “Deep Head Pose Estimation from Depth Data for In-Car Automotive Applications”. In: *Understanding Human Activities Through 3D Sensors*. 2018, pp. 74–85.

[132] Y. Wang, Y. Ren, S. Elliott, and W. Zhang. “Enabling Courteous Vehicle Interactions through Game-based and Dynamics-aware Intent Inference”. In: *IEEE Trans. on Intelligent Vehicles (TIV)* 5.2 (2020), pp. 217–228.

[133] Y. Wang, W. Liang, J. Shen, Y. Jia, and L.-F. Yu. “A deep Coarse-to-Fine network for head pose estimation from synthetic data”. In: *Pattern Recognition* 94 (2019), pp. 196–206.

[134] G. Welch and G. Bishop. “An Introduction to the Kalman Filter”. In: *In Practice* 7.1 (2006), pp. 1–16.

[135] WHO. “Global Status Report on Road Safety”. In: *World Health Organization* (2018).

[136] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays. “Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting”. In: *Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*. 2021.

[137] H. Yan, W. Doerr, A. Ioffe, and H. Clasen. “Micro-Doppler Based Classifying Features for Automotive Radar VRU Target Classification”. In: *Int. Tech. Conf. on the Enhanced Safety of Vehicles (ESV)*. 2017, pp. 1–8.

[138] Y. Yan, Y. Mao, and B. Li. “SECOND: Sparsely Embedded Convolutional Detection”. In: *Sensors* 18.10 (2018), p. 3337.

[139] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang. “FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation From a Single Image”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1087–1096.

[140] Y. Yu, K. Funes Mora, and J.-M. Odobez. “HeadFusion: 360° Head Pose tracking combining 3D Morphable Model and 3D Reconstruction”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 40.11 (2018), pp. 2653–2667.

[141] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda. “A Survey of Autonomous Driving: Common Practices and Emerging Technologies”. In: *IEEE Access* 8.1 (2020), pp. 58443–58469.

[142] W. Zeng, S. Wang, R. Liao, Y. Chen, B. Yang, and R. Urtasun. “DSDNet: Deep structured self-driving network”. In: *European Conf. on Computer Vision (ECCV)*. 2020, pp. 156–172.

[143] Z. Zhang. “A flexible new technique for camera calibration”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 22.11 (2000), pp. 1330–1334.

[144] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. “On the Continuity of Rotation Representations in Neural Networks”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 5738–5746.

[145] Y. Zhou and J. Gregson. “WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose”. In: *Proc. of the British Machine Vision Conference (BMVC)*. 2020, pp. 1–13.

[146] Y. Zhou and O. Tuzel. “VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 4490–4499.

PROPOSITIONS

accompanying the dissertation

DRIVER AND PEDESTRIAN MUTUAL AWARENESS FOR PATH PREDICTION IN INTELLIGENT VEHICLES

by

Markus ROTH

1. Using the head pose distribution of naturalistic driving leads to inferior head pose estimation systems.
   This proposition pertains to Chapter 4.

2. The choice of head pose representation has a significant influence on the complexity of the developed head pose estimation method architecture.
   This proposition pertains to Chapter 4.

3. Future path prediction systems will outperform current systems by learning superior context cues from data. This comes at the cost of lack of interpretability.
   This proposition pertains to Chapter 5.

4. A 3D pose cannot be judged sufficiently by observing its projections onto a camera image.

5. The effort of dataset collection and curation increases super-linear with the number of sensors in-use.

6. "Driver" observation is also relevant for self-driving vehicles.

7. Sharing data among the automotive industry will considerably speed up the development of automated driving.

8. Pedestrians who understand the pedestrian motion models employed in intelligent vehicles will take advantage of them.

9. The principles of planning apply equally to research. This includes planning fallacy.

10. Writing a thesis is easy: just backpropagate the error from the doctoral regulations.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotor prof. dr. D.M. Gavrila and the copromotor dr. J.F.P. Kooij.