On-road Assessment of Driver Workload and Awareness in Automated Vehicles

Stapel, J.C.J.

DOI
10.4233/uuid:746f5f73-1876-4371-b142-f0f3117ded6a

Publication date
2021

Document Version
Final published version

Citation (APA)
Stapel, J. C. J. (2021). *On-road Assessment of Driver Workload and Awareness in Automated Vehicles*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:746f5f73-1876-4371-b142-f0f3117ded6a

ON-ROAD ASSESSMENT OF DRIVER WORKLOAD AND AWARENESS IN AUTOMATED VEHICLES

Jaap Cornelis Jork Stapel

On-road Assessment of Driver Workload and Awareness in Automated Vehicles

On-road Assessment of Driver Workload and Awareness in Automated 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
Thursday, 25 February 2021 at 15:00 o'clock

by

Jaap Cornelis Jork Stapel

Master of Science in Aerospace Engineering, Delft University of Technology,
Delft, The Netherlands
born in Haarlem, The Netherlands.

This dissertation has been approved by the promotors:

Dr. ir. R. Happee and Prof. dr. D.M. Gavrila

Composition of the doctoral committee:

Rector Magnificus,	Delft University of Technology, chairperson
Dr. ir. R. Happee,	Delft University of Technology, promotor, 3mE
Prof. dr. D.M. Gavrila,	Delft University of Technology, promotor, 3mE

Independent members:

Prof. P.R. de Lucia	Rice University, USA
Prof. dr. ir. C.N. van Nes	Delft University of Technology
Prof. dr. D. de Waard	Rijksuniversiteit Groningen
Prof. dr. ir. B. van Arem	Delft University of Technology
Dr. ir. J.C.F. de Winter	Delft University of Technology

This work was supported by the NWO-TTW Foundation, the Netherlands, under the project "From Individual Automated Vehicles to Cooperative Traffic Management - Predicting the benefits of automated driving through on-road human behavior assessment and traffic flow models (IAVTRM)" -STW# 13712.

Keywords: situation awareness, naturalistic driving, driver support

Printed by: Gildeprint

Cover by: Jip van Montfort

ISBN 978-94-6419-134-9

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

If you can't do it the way you should,
you should do it the way you can.

Grandma

Contents

SUMMARY

PROBLEM DEFINITION

According to the World Health Organization, traffic injuries have become the eighth cause of death and the leading cause among children and young adults. Human error, and in particular perceptual error, is among the most frequently reported causes of road fatalities. The desire to reduce traffic fatalities has led to the development of automated driving, which promises revolutionary advances in driver safety, traffic capacity and driver convenience. Since true autonomy in mixed traffic has not yet been achieved, today's automated vehicles require the driver to continuously supervise the automation and to capably intervene when necessary. However, simulator studies and experiences from disciplines such as aviation and factories have demonstrated that humans are generally ill-equipped to monitor automation for longer periods. This raises the concern that partial automation may harm rather than help traffic safety if not designed to adequately support the drivers in their supervisory tasks.

RESEARCH OBJECTIVES

To address this concern, further insights are needed in how drivers monitor automation in complex real-world traffic, and how their behaviour and performance change with long-term automated driving experience. This dissertation sets out to investigate how real-world automation changes the availability of attentional resources, to establish where and how drivers use automation in naturalistic conditions, and evaluate how these change with experience. While these objectives investigate periods of automated driving, vehicles with automated driving functionalities will often be driven manually, when outside the operational design domain or at the driver's preference. In these conditions, the available automation may still outperform the driver on particular tasks, such as detecting and tracking surrounding road users without bias or distraction. This dissertation therefore also contributes to the search for ways in which automation can provide meaningful support to the traffic monitoring task in manual and supervised driving.

To evaluate if and when supervised automated driving negatively affects the driver's ability to monitor, mental workload is evaluated in a Tesla model S on public roads (Chapter 2). Voluntary automation use and attention are examined in a naturalistic driving study on public roads (Chapter 3). To evaluate the effect of experience with automated driving, Chapter 2 compares drivers with and without prior automation use, whereas Chapter 3 examines how behaviour changes over a two-month period, compared to one month of manual driving.

Two studies are performed to examine how driving automation can support the driver with the monitoring task, for which an instrumented vehicle was extended with cameras which track the driver's gaze and associate it to surrounding road users as detected by the vehicle perception. The first study (Chapter 4) investigates how well gaze behaviour can indicate driver awareness toward individual road users, and proposes a recognition task to obtain

a ground truth for awareness of multiple other road-users. The second study (Chapter 5) evaluates if driver gaze and head pose can provide earlier predictions for emergency alerting and intervention systems. A crossing pedestrian collision risk prediction system is used as a case study where gaze and contextual cues are evaluated in their contribution to path and risk prediction using a dynamic Bayesian network.

FINDINGS & RECOMMENDATIONS

Chapter 2 found that workload differed between roads with high and low traffic complexity, both for manual and automated driving, which indicates that drivers remain sensitive to changes in task demand while supervising automated driving. Drivers with prior experience in automated driving perceived a lower workload while supervising automation compared to manual driving. No workload difference was perceived for first-time users. In contrast, attentional demand as measured by a detection-response task was higher during automation use compared to manual driving regardless of experience. This indicates that monitoring automation (SAE2) requires more mental capacity compared to manual driving, which suggests that in contrast to a wide range of studies, SAE2 can increase workload. Supervising automation may therefore be beneficial for driver attention, but perception of workload during supervision may be too low for this to occur naturally. Future work should consider calibrating workload perception and system limitation understanding rather than actual task demand to encourage attentive supervision.

Chapter 3 shows that automation is mostly used on road types generally considered suitable for automated driving with only incidental use on urban roads. This suggests that users are adhering to the operational design domain of these vehicles. On highways, automation is used at all speeds, but less during short periods of slow driving. No time-in-drive, time-of-day or experience effects were found for automation use. On the highway, head pose deviation was smaller during automation use compared to manual driving but tended to increase over the first six weeks of use, which may indicate a change in monitoring strategy. Further research is needed to assess if this difference indicates better or worse monitoring behaviour.

Chapter 4 found that drivers performed better on the recognition task when road users were relevant for the driven manoeuvre and when drivers had directed their gaze within 10 degrees of these road users. However, at least 18% of road users were recognised while only observed peripherally, suggesting that peripheral vision should not be neglected in attention monitoring. Recognition performance was not predicted by gaze metrics and requires further development to reduce forget rates. Further analysis is needed to compare the recognition task to established situation awareness measures after these improvements are obtained.

Chapter 5 demonstrates that driver and pedestrian attention monitoring can provide a benefit to pedestrian crossing collision risk prediction when predicting further than 0.75 seconds ahead.

The higher workload during supervised automation and the general adherence to the operational design domain in naturalistic driving indicate that supervising driving automation can be beneficial to driver attention and traffic safety, but literature and recent accidents demonstrate that challenges remain in encouraging such attentive behaviour. Strategies to encourage attentive supervision should therefore be further developed, as well as ways to maintain these strategies while automation technology improves in pursuit of the opposite objective to reduce engagement in the driving task.

The joint analysis of driver gaze and road scene may improve driver support during manual driving and supervised automation, and benefit the development of automated driving. But care should be taken that systems which use driver attention or rely on other contextual cues do not become susceptible to the same mistakes as drivers tend to make. While careful design approaches can reduce the risk of mimicking human error, validation will ultimately require a reliable way to distinguish between awareness and inattentional blindness.

The instrumentation and conducted studies with on-road automation demonstrate that on-road research is becoming more practical and accessible than ever before, thanks to recent developments in automation. The observation that during on-road automation, inexperienced drivers perceive higher workload compared to in simulators testifies for the importance of on-road driving research. Challenges encountered during the naturalistic study and attention study demonstrate that the instrumentation and processing have to be designed and tested carefully for on-road research to be effective.

1

Introduction

1.1. PROBLEM DEFINITION

According to the World Health Organization, traffic injuries are now the eighth cause of death and the leading cause among children and young adults (WHO, 2018). While collision mitigation and driver assistance systems have provided major safety improvements over the past few decades, road fatality rates per unit of population have stayed constant between 2013 and 2017 for most world regions. Human error, and in particular perceptual error, is among the most frequently reported causes of road fatalities (European Road Safety Observatory, 2018).

This has advocated a strong incentive for the recent developments towards automated driving, with prospects of revolutionary advances in driver safety, traffic capacity and driver convenience. Since true vehicle autonomy within today's traffic has proven to be an enormous challenge, developments are released to the public incrementally through various forms of partial automation, where the driver remains responsible to ensure safety.

In 2014, the automated driving committee of the international society of automotive engineers (SAE) produced a widely adopted standard for classifying the many forms of vehicle automation, and revised it over consecutive years (SAE International, 2018). This taxonomy divides automation in five levels according to the distribution of responsibilities between driver and automation, as shown in Figure 1.1.

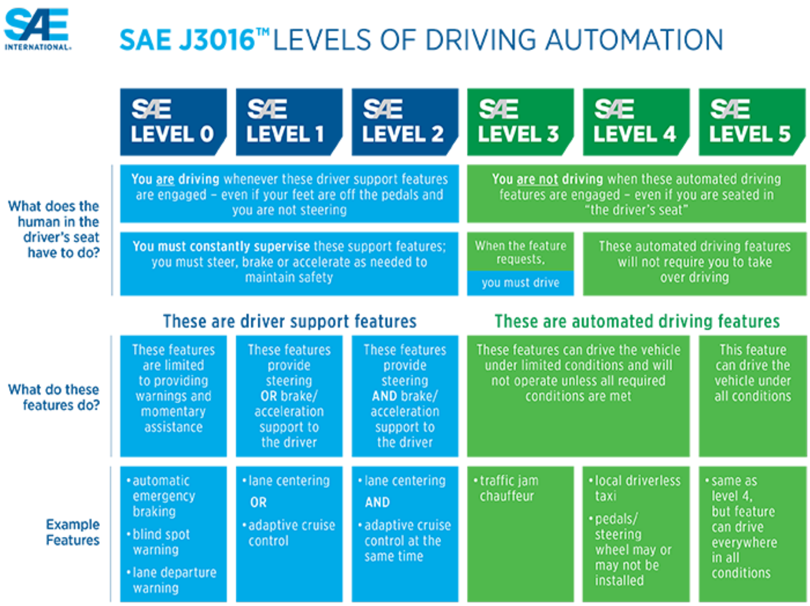


Figure 1.1: SAE levels of driving automation. Adopted from SAE International (2019)

Until the automation can flawlessly interpret all the complexities of the traffic environment and our behaviours within it (SAE4/5), human drivers will have to complement these systems. In SAE2 systems, the driver is required to continuously monitor the automation, and to capably intervene when necessary. In SAE3, the driver may engage in other tasks but should be able to resume driving on short notice when requested by the automation. Since automation of levels SAE0 to SAE4 only functions reliably under specific conditions (e.g. highway driving, congested traffic, flat road sections with clear lane markings), the level of support a given vehicle can give may vary with time and location. While automated driving can only responsibly control the vehicle within the operational design domain (ODD), it may still play a supportive role while the driver operates the car manually. Such support comes as SAE0 active safety features which co-monitor the environment and either inform the driver of possible hazards (e.g. forward collision warning, blind spot monitoring) or intervene when it becomes evident that the driver will not react to an imminent risk (e.g. automated emergency braking). These systems have been found to make a positive impact on driving safety statistics (Cicchino, 2017, 2018). In contrast, continuous automation of speed regulation and lane keeping as implemented in current levels SAE1 and SAE2 have not yet provided clear improvements in accident statistics and raised concerns of unsafe use of automation (Dijsselbloem *et al.*, 2019; Vlakveld, 2019). Concerns are raised by researchers in the field of human factors which caution that "autopia" may contain dystopian elements where (driving) automation may harm rather than help traffic safety and quality of life in general (Hancock, 2019). This dissertation addresses immediate safety implications in the interaction between driver and vehicle.

Despite the clear definition of roles, supervised automation (SAE 2/3) inevitably forms a transition between support and autonomy, which can cause confusion about the user's expectations and beliefs (Victor *et al.*, 2018). Drivers may not be fully aware of their vehicle's capabilities (Harms *et al.*, 2020), possibly leading to inattention, unsafe use or under-use. Zhang *et al.* (2019) review that several determinants of reaction time to take over requests are voluntary (secondary tasks, usage of available take over time), demonstrating that attentive supervision is often challenged by motivational rather than mental or physical limitations. These reaction times increase as drivers get more engaged in secondary tasks (Ko and Ji, 2018).

Even when these issues are addressed through clear communication and reminders, supervised automation still changes the driver's role, which introduces further challenges. Skill degradation from reduced practice in manual driving has been identified as a concern (Miller and Boyle, 2018), and frequent transitions in responsibilities may lead to mode confusion and schema-type slips (Norman, 1981). The transition from continuous driving to passive monitoring also challenges driver's vigilance and takeover performance.

Furthermore, automation reduces perceived workload and increases productivity on secondary tasks (de Winter *et al.*, 2014). While these effects seem advantageous from a comfort perspective, they also raise safety concerns. When the workload gets too low, mental underload may occur (de Waard, 1996), making it harder to pay attention to the task. Over time, this can lead to a state of drowsiness (Vogelpohl *et al.*, 2018), inattention and slower reactions (Greenlee *et al.*, 2018). In supervised automation, this form of drowsiness can develop within 15 minutes of monotonic driving (Goncalves *et al.*, 2016), and already in manual driving, the development of drowsiness forms a contributing factor in 20% of road accidents (MacLean *et al.*, 2003).

These risks of supervised automation (engagement in secondary tasks, schema type or habitual errors and impaired mental state) can all contribute to perceptual error, and impair the driver's situation awareness. While workload and driver state can impair our ability to monitor, driver behaviour may moderate these effects. Risk compensation and misuse may lead to insufficient situation awareness or time to respond. This may be reduced by compensatory strategies, such as strategic planning of breaks, secondary tasks and automation usage, which may be adopted naturally but can also be actively encouraged through system and HMI design or driver training. However, human factors research generally concludes that these countermeasures cannot fully overcome the reduced energetic state which automation imposes on the driver, and that further improvements in automation reliability will make such countermeasures less effective because without the perception of risk or error, there is no intrinsic motivation to monitor the automation (Hancock, 2013). Other studies suggest that the mental model of the automation's operational design domain can be incorrect in many circumstances of automation experienced and -inexperienced drivers alike (Farah *et al.*, 2020). However, there are also indications that experience with automation and automation failure can improve the driver's ability to intervene (Zhang *et al.*, 2019). Alternatively, automation can be altered to provide support for tasks which are challenging for the driver (e.g. preventing attentional lapses), while limiting the extent to which easy tasks are automated (Cabrall *et al.*, 2019; Mulder *et al.*, 2012).

1.2. OBJECTIVES

Many excellent simulator and test track studies have contributed to our understanding of the driver's interaction with automation (de Winter *et al.*, 2014; Parnell *et al.*, 2018; Zhang *et al.*, 2019). To prepare society for automated driving, it is imperative that we extend this knowledge with empirically founded insights in how drivers perform and behave in complex real-world automation, and to which extent concernable behaviours affect overall safety (Banks *et al.*, 2018; Fridman *et al.*, 2018; Jarosch *et al.*, 2019; Naujoks *et al.*, 2016). More attention is needed to examine behaviour and performance of drivers with long-term experience of today's (but also tomorrow's) driving automation. Particularly, there is a paucity in studies that include drivers with long-term, real-world experience with driving automation.

Since perceptual errors are amongst the leading factors in fatal crashes and since automation has been demonstrated to negatively impact monitoring ability, I want to better understand what makes (automated) driving hard or easy to attend to, and want to establish how monitoring is affected by experience, and if this results in better or worse automation monitoring. I set out to evaluate how current SAE2 automation affects the driver's attentional state in the real world, and how this varies with driving conditions and driver experience.

Since the impact of automation on the driver depends on how it is used, I also want to better understand how different driving situations influence automation use in naturalistic conditions, whether drivers adopt different monitoring strategies under different circumstances, and how these strategies evolve as the drivers gain more experience in using these systems.

While the preceding objectives investigate periods of automated driving, vehicles with these functionalities will often be driven manually, when outside the ODD or at the driver's preference. In these conditions, the available automation may still outperform the driver on particular tasks, such as detecting and tracking surrounding road users without bias or distraction. I therefore also contribute to the search for ways in which these automation components can provide meaningful support to the monitoring task during manual (and possibly during SAE2) driving.

On-road research imposes additional challenges to the measurement and control of what transpires around the vehicle. While the importance of high-fidelity research and validation of simulator studies is widely accepted, there are two concerns which tend to make on-road research a relatively unpopular choice (de Winter *et al.*, 2012): the lack of experimental control in on-road studies makes it hard to reduce variation in the independent and control variables, and to control for confounding factors. Secondly, there is more effort involved in the collection of dependent variables such as what happens to and around the vehicle. With the present developments of automated driving, this last challenge is nearly resolved since considerable effort is invested in reliable sensing and automated interpretation of the vehicle surroundings. The lack of control under real-world experiments makes on-road research especially unpopular for hypothesis testing, because the increased variance makes it harder to demonstrate effects statistically. However, driver research is closer to the applied than to the fundamental end of the scale, and the envisioned applications for driver monitoring and driver support will ultimately have to cope with the diversity and confounders of real-world driving. Familiarity with these real-world complications will benefit researchers who hope to develop systems and tools that will meet the interned prospects once implemented, and exclusively practicing research in idealized conditions designed to elicit maximum effects will not provide such insights.

This dissertation fully pursues this by avoiding the simulator entirely and conducting all research on-road. It also aims to illustrate that on-road research is becoming more practical and accessible than ever before, thanks to recent developments in automation.

Figure 1.2: An instrumented vehicle used for this research, equipped with eye tracking and computer vision. The green line and marker represent the visual focus detected using eye-tracking and is not visible while driving.

This leads to the following research objectives and questions:

1. Quantify the effect of real-world automation on the availability of attentional resources
   (a) How does cognitive load differ between manual and automated driving?
   (b) How does traffic complexity affect cognitive load?

1

   (c) How does automation experience affect cognitive load?
   (d) What are the implications for monitoring quality and safety?

2. Gain insight in where and how drivers use automation, and how this changes with experience.
   (a) Which conditions affect when driving automation is used?
   (b) To which extent is automation used outside the (safe) design domain?
   (c) How does automation use affect attention behaviour?
3. Can driving automation technology support the driver's monitoring task by interpreting the driver's awareness of individual cues in the road scene?
   (a) Can a recognition task be used to evaluate awareness toward individual road users?
   (b) Can situation awareness be predicted from gaze metrics relative to individual road users?
   (c) Can traffic awareness monitoring be used to provide a prediction benefit collision warning and avoidance systems?

1.3. APPROACH

The first two objectives are pursued using current SAE2 automation using a controlled experiment on public roads (Chapter 2), and a naturalistic driving study on public roads (Chapter 3). The third objective is pursued using our instrumental automated vehicle in staged experiments (Chapters 4 and 5).

To evaluate under which conditions automated driving may negatively affect the driver's ability to monitor, Chapter 2 examines how automation affects the driver's cognitive workload. To make a within-subject comparison of cognitive workload during attentive manual and automated driving, an on-road experiment was performed in a Tesla Model S. Subjective, psychological and performance indicators of cognitive workload were compared between manual and automated driving, under the influence of two moderating factors: automation experience (comparing driving automation-novices to Tesla owners) and traffic complexity (comparing a quiet highway to the round-way of Amsterdam).

While Chapter 2 provides insights in driver workload during automation use, the participants are instructed to show normative, attentive behaviour. Real-world behaviour is likely to differ; drivers may choose when to use the automation and how to divide their attention between monitoring the road and competing secondary tasks like texting on a phone. This implies that the idealised capabilities obtained in the workload study only capture a partial image of the safety impact of automated driving. Behavioural adaptation to automated driving is best examined in a naturalistic study, where participants drive their vehicles in their daily life without being influenced by instructions or the presence of an experimenter. Naturalistic studies however result in large amounts of data whose condensation into knowledge can be demanding. Chapter 3 contributes to the data enrichment and analysis of a naturalistic driving study aimed to examine behavioural adaptation to the introduction of SAE2-capable vehicles. The study is unique in its inclusion of a baseline period without automation use, which allows for a within-subject comparison of behavioural adaptation.

Chapter 4 and Chapter 5 investigate how vehicle perception of the environment can support the driver in monitoring the environment. Meaningful attention support poses two challenges: it requires the automation system to 1) understand which road users require the driver's awareness, and 2) measure which of these the driver is aware of, and which are overlooked. To address the latter, Chapter 4 investigates how well gaze behaviour can indicate driver awareness toward individual road users during left turns on complex urban intersections using a vehicle with road-scene perception and eye tracking. This evaluation also requires a ground truth for this awareness to be obtained for multiple road users simultaneously without burdening the driver with additional tasks. A recognition-based method is developed for labelling driver situation awareness.

Chapter 5 evaluates whether attention monitoring can provide a temporal advantage to emergency alerting and intervention systems. A crossing pedestrian collision risk prediction system is used as a case study. In order to be anticipative instead of reactive, the system has to not only predict the travelled paths of the pedestrian or driver, but also reliably predict if they intend to adapt their behaviour to resolve a potential collision. For this purpose, a dynamic Bayesian network (DBN) is developed to predict the attentional and intentional hidden states of the driver and pedestrian, as well as other contextual cues. The prediction performance is compared to an awareness agnostic system, as well as the availability of other candidate context observations.

1.4. DISSERTATION CONTRIBUTIONS

Chapter 2 examined how supervising SAE2 automation affects perceived workload and attentional demand on two highways of different traffic complexity with automation experienced and -inexperienced drivers. The results show that automation experience and traffic complexity have substantial effects on workload while monitoring SAE2 automation. Perceived workload and objective workload show the same trend while comparing driving environments; driver workload remained sensitive to changes in traffic complexity during supervised automated driving, which implies that drivers maintain the ability to mobilize attentional resources depending on the situation's need. However, the objective and subjective workload measures show opposite effects when examining the effect of automation, in particular for automation-experienced drivers. Automation-experienced drivers perceived a lower workload during automation use. However, their performance on the objective workload task indicates that monitoring SAE2 automation requires more mental capacity compared to manual driving, which suggests that in contrast to a wide range of studies, SAE2 can increase workload. Hence, SAE2 can alleviate unsafe mental under-load rather than cause it. Our on-road testing further identifies a workload difference between automation-experienced and automation-inexperienced drivers which was not observed in reviewed simulator studies, highlighting the importance of testing automation experienced users.

In Chapter 3, a naturalistic dataset is enriched and explored to examine automation usage and attention during the first two months of using SAE2-capable vehicles. For data enrichment, neural networks were trained to classify automation status and the driver's direction of attention. For status classification, template matching of instrument icons was used and attention was classified from video-derived head pose. While status estimation was found to be reliable, the second network was unable to distinguish between attentive and distracted regions of attention, despite performing on-par with reviewed methods. This shows a need for eye tracking rather than using head pose. Results of the automation status classifier were used to explore when and in which driving conditions SAE2 automation is used, and if these patterns change with experience. Automation is mostly used on road types generally considered suitable with only incidental use on urban roads. This suggests that users are adhering to the operational design domain of these vehicles. On highways, automation is used at all speeds, but less during short periods of slow driving. No differences in usage were observed for time in trip, time of day or experience. During highway automation use, head pose deviation did not differ between SAE2 automation and baseline manual driving, but tended to increase over the first six weeks of use, which hints at behavioural adaptation. Head heading and pitch deviation were smallest during ACC use. Further research is needed to assess if this difference indicates better or worse monitoring behaviour.

Chapter 4 examines if driving automation technology can interpret the driver's awareness towards individual road users. Driver gaze is associated with surrounding road users as detected by computer vision during left turns on urban intersections. A post-drive recognition task was performed to assess driver awareness. Typical gaze behaviour towards various road users during left turn manoeuvres could predict road user relevance but not the outcome of the recognition task. The recognition task was sensitive to road user relevance and minimum gaze angle, and yielded a low false positive rate, which demonstrates it can identify awareness of individual road users during left turn manoeuvres. However, the true positive rate was unexpectedly low, for which solutions were proposed. The findings further show that perception occurs at gaze angles well beyond 10° which suggests that perception models should incorporate more than fixation location in their parameterization.

Chapter 5 demonstrates that contextual cues including driver and pedestrian awareness through gaze and head pose provide a temporal benefit on collision risk prediction performance for prediction horizons beyond 1.5 s. The findings also show that collision course is an insufficient cue to disambiguate who will yield when both driver and pedestrian are attentive. Additional cues such as mutual awareness (i.e. knowing the other's awareness of oneself) and knowledge on right of way are required to correctly predict who may yield.

1.5. REFERENCES

V. A. Banks, A. Eriksson, J. O'Donoghue, and N. A. Stanton, *Is partially automated driving a bad idea? observations from an on-road study,* Applied ergonomics **68**, pp. 138–145 (2018).

C. D. D. Cabrall, A. Eriksson, F. Dreger, R. Happee, and J. de Winter, *How to keep drivers engaged while supervising driving automation? a literature survey and categorisation of six solution areas,* Theoretical Issues in Ergonomics Science **20** (3), pp. 332–365 (2019).

J. B. Cicchino, *Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates,* Accident analysis and prevention **99** (Pt A), pp. 142–152 (2017).

J. B. Cicchino, *Effects of blind spot monitoring systems on police-reported lane-change crashes,* Traffic injury prevention **19** (6), pp. 615–622 (2018).

J. Dijsselbloem, M. van Asselt, S. Zouridis, and C. Verheij, *Wie stuurt? verkeersveiligheid en automatisering in het wegverkeer,* (2019).

European Road Safety Observatory, *Traffic safety basic facts on junctions,* (2018).

H. Farah, S. Bhusari, P. van Gent, F. A. M. Babu, P. Morsink, R. Happee, and B. van Arem, *An empirical analysis to assess the operational design domain of lane keeping system equipped vehicles combining objective and subjective risk measures,* IEEE Transactions on Intelligent Transportation Systems , pp. 1–10 (2020).

L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, *Cognitive load estimation in the wild,* CHI Conference on Human Factors in Computing Systems , pp. 1–9 (2018).

J. Goncalves, R. Happee, and K. Bengler, *Drowsiness in conditional automation: Proneness, diagnosis and driving performance effects,* in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (IEEE, 2016) pp. 873–878.

E. T. Greenlee, P. R. DeLucia, and D. C. Newton, *Driver vigilance in automated vehicles: Hazard detection failures are a matter of time,* Human Factors and Ergonomics Society , pp. 1–12 (2018).

P. A. Hancock, *In search of vigilance: The problem of iatrogenically created psychological phenomena,* American Psychologist **68** (2), pp. 97–109 (2013).

P. A. Hancock, *Some pitfalls in the promises of automated and autonomous vehicles,* Ergonomics **62** (4), pp. 479–495 (2019).

I. M. Harms, L. Bingen, and J. Steffens, *Addressing the awareness gap: A combined survey and vehicle registration analysis to assess car owners' usage of adas in fleets,* Transportation Research Part A: Policy and Practice **134**, pp. 65–77 (2020).

O. Jarosch, S. Paradies, D. Feiner, and K. Bengler, *Effects of non-driving related tasks in prolonged conditional automated driving – a wizard of oz on-road approach in real traffic environment,* Transportation Research Part F: Traffic Psychology and Behaviour **65**, pp. 292–305 (2019).

S. M. Ko and Y. G. Ji, *How we can measure the non-driving-task engagement in automated driving: Comparing flow experience and workload,* Applied ergonomics **67**, pp. 237–245 (2018).

A. W. MacLean, D. R. Davies, and K. Thiele, *The hazards and prevention of driving while sleepy,* Sleep Medicine Reviews **7** (6), pp. 507–521 (2003).

E. E. Miller and L. N. Boyle, *Behavioral adaptations to lane keeping systems: Effects of exposure and withdrawal,* Human factors , pp. 152–164 (2018).

M. Mulder, D. A. Abbink, and E. R. Boer, *Sharing control with haptics: Seamless driver support from manual to automatic control,* Human factors **54** (5), pp. 786–798 (2012).

1

F. Naujoks, C. Purucker, and A. Neukum, *Secondary task engagement and vehicle automation – comparing the effects of different automation levels in an on-road experiment,* Transportation Research Part F: Traffic Psychology and Behaviour **38**, pp. 67–82 (2016).

D. de Waard, *The measurement of drivers' mental workload,* Ph.D. thesis, University of Groningen (1996).

J. C. F. de Winter, R. Happee, M. H. Martens, and N. A. Stanton, *Effects of adaptive cruise control and highly automated driving on workload and situation awareness, a review of the empirical evidence,* Transportation Research Part F: Traffic Psychology and Behaviour **27** (B), pp. 196–217 (2014).

J. C. F. de Winter, P. M. van Leeuwen, and R. Happee, *Advantages and disadvantages of driving simulators: A discussion,* Proceedings of Measuring Behavior Conference (2012).

D. A. Norman, *Categorization of action slips,* Psychological Review **88** (1), pp. 1–15 (1981).

K. J. Parnell, N. A. Stanton, and K. L. Plant, *Good intentions: Drivers' decisions to engage with technology on the road and in a driving simulator,* Cognition, Technology & Work **26** (4), p. 597–619 (2018).

SAE International, *J3016 jun2018: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles,* (2018).

SAE International, *Sae international releases updated visual chart for its "levels of driving automation" standard for self driving vehicles.* (2019).

T. W. Victor, E. Tivesten, P. Gustavsson, J. Johansson, F. Sangberg, and M. Ljung Aust, *Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel,* Human factors **60** (8), p. 1095–1116 (2018).

W. P. Vlakveld, *Veiligheidseffecten van rijtaakondersteunende systemen: Bijlage bij het convenant van de adas alliantie,* (2019).

T. Vogelpohl, M. Kühn, T. Hummel, and M. Vollrath, *Asleep at the automated wheel-sleepiness and fatigue during highly automated driving,* Accident analysis and prevention **126** (2018), 10.1016/j.aap.2018.03.013.

WHO, *Global status report on road safety 2018: Summary,* World Health Organization (WHO/NMH/NVI/18.20) (2018).

B. Zhang, J. de Winter, S. Varotto, R. Happee, and M. Martens, *Determinants of take-over time from automated driving: A meta-analysis of 129 studies,* Transportation Research Part F: Traffic Psychology and Behaviour **64**, pp. 285–307 (2019).

2

AUTOMATED DRIVING REDUCES PERCEIVED WORKLOAD, BUT MONITORING CAUSES HIGHER COGNITIVE LOAD THAN MANUAL DRIVING

Jork Stapel, Freddy Mullakkal-Babu, Riender Happee

This chapter has been published as: Jork Stapel, Freddy Antony Mullakkal-Babu, Riender Happee, *Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving*, Transportation research part F: traffic psychology and behaviour, 60 (2019).

Abstract

Driver mental workload is an important factor in the operational safety of automated driving. In this study, workload was evaluated subjectively (NASA R-TLX) and objectively (auditory detection-response task) on Dutch public highways (~150 km) comparing manual and supervised automated driving in a Tesla Model S with moderators automation experience and traffic complexity. Participants (N=16) were either automation-inexperienced drivers or automation-experienced Tesla owners. Complexity ranged from an engaging environment with a road geometry stimulating continuous traffic interaction, and a monotonic environment with lower traffic density and a simple road geometry. Perceived and objective workload increased with traffic complexity. When using the automation, automation-experienced drivers perceived a lower workload, while automation-inexperienced drivers perceived their workload to be similar to manual driving. However, the detection-response task indicated an increase in cognitive load with automation, in particular in complex traffic. This indicates that drivers under-estimate the actual task load of attentive monitoring. The findings also highlight the relevance of using system-experienced participants and the importance of incorporating both objective and subjective measures when examining workload.

2.1. INTRODUCTION

Monitoring ability is essential in an increasing number of vehicles offering supervised, or SAE2 automation (SAE International, 2016), which require the driver to monitor the automation and intervene when needed. Driver mental workload is an important factor in the operational safety of supervised automation. When automation relieves the driver from the continuous control tasks, mental underload can occur (de Waard, 1996). Over time, this can lead to a state of drowsiness, inattention and slower reactions (Greenlee, DeLucia, & Newton, 2018; Hirose, Kitabayashi, & Kubota, 2015). This has raised concerns regarding the driver's ability to monitor the automation and his/her performance to intervene in critical situations (Kyriakidis et al., 2017).

In order to address these effects, it is important to know how workload is affected by the use of automation, and how this effect varies with driving conditions. This study focuses on two main moderating variables of workload: the complexity of the driving environment and the driver's experience with the automation. Understanding the effect of these moderators can help to predict in which conditions workload is too high or too low. Experience with driving automation can lead to task execution at a lower cognitive level, or reduce the perceived complexity of the traffic situation (Paxion, Galy, & Berthelon, 2014; Young & Stanton, 2007). Automation experience can also lead to better monitoring and improved cognitive readiness for familiar driving situations, resulting in higher control transition performance (Krampell, 2016; Larsson, Kircher, & Andersson Hultgren, 2014; Paxion et al., 2014; Wright, Samuel, Borowsky, Zilberstein, & Fisher, 2016; Young & Stanton, 2007). Moreover, automation experience may reduce task demand, or reduce sensitivity to demand changes, and thus influence workload differently in high and low traffic complexity (Patten, Kircher, Ostlund, Nilsson, & Svenson, 2006; Stanton, Hedge, Brookhuis, Salas, & Hendrik, 2005).

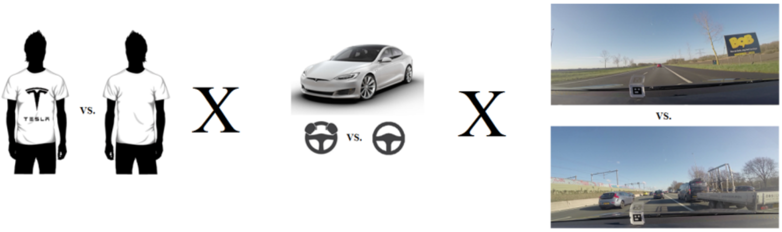


Figure 2.1: Illustration of the independent variables: automation experience, automation use, and complexity of the environment.

This study investigates how workload changes with monitored automated driving in real-world conditions, and how this change is moderated by traffic complexity and by the driver's prior experience with automated driving. We conducted an on-road experiment on Dutch public highways in a Tesla Model S. The change in workload was assessed subjectively (NASA R-TLX) as well as objectively (auditory detection-response task). Traffic complexity was moderated by driving in a monotonic, low workload and a complex, engaging highway. To moderate automation experience, participants were either automation-inexperienced drivers or automation-experienced Tesla owners. The conditions were driven both manually and with automation. This resulted in a 2 (automation: on vs. off) x 2 (environment: monotonic vs. engaging) x 2 (experience: experienced vs. inexperienced) mixed design as illustrated in Figure 2.1.

2.1.1. THEORIES OF WORKLOAD

In line with resource theory and the capacity model (Kahneman, 1973), we describe workload as the ratio between task demands and resources available to meet them. (A discussion of alternative definitions can be found in (Cain, 2007)). Task demand depends on the complexity of the driving task and the traffic situation, but also on how the goals are set (i.e. accepting a level of performance), and the strategy chosen to achieve it. To meet these demands, the driver has to allocate physical and mental resources, which are limited in availability. Driving consists of multiple sub-tasks. To model when and how much these tasks interfere, Wickens (1981) proposed the multiple resource theory in which resource pools are available for the different modalities of perception (e.g. visual, auditory, tactile), the codes of processing (spatial or verbal) and response selection and execution (hands, feet, speech). In addition, he proposed a cognitive resource shared across all tasks.

Resources are finite in capacity, but the upper limit is considered elastic (Kahneman, 1973; Young & Stanton, 2002), and closely related to the driver's energetic state. Drivers may exert state related effort to improve their energetic state. Investing computational effort can compensate for increasing demand. Both forms of effort are consciously perceived, and are considered key aspects of perceived workload (de Waard, 1996).

The relation between task demand and workload is u-shaped (de Waard, 1996) and consists of regions of underload, optimal load and overload. In optimal load, performance is generally good and changes in demand have little or no effect on perceived effort or achieved performance. Overload occurs when demands exceed the available resource capacity and performance degrades despite the additional effort invested. Underload occurs when demands are exceptionally low or monotonous in nature. Underload can lead to vigilance decrement, or inattention. However, low task demand can lead to an increase in workload when drivers recognize the development of drowsiness and invest state-related effort to compensate (Warm, Parasuraman, & Matthews, 2008).

Experience can make some demanding tasks impose less or no effort, even when performed concurrently with effortful tasks. These include routine operations and learned skills, executed with a high degree of automaticity. Examples are lane keeping, speed or headway maintenance and event detection. When automatized routines can handle the situation, these driving tasks should be insensitive to changes in cognitive load. According to the cognitive control hypothesis, cognitive load from competing tasks can only emerge for non-automatized tasks or when overruling skill-based behavior (Engström, Markkula, Victor, & Merat, 2017). We thus expect automation-experienced drivers to have a lower workload during automation compared to automation-inexperienced drivers. Conversely, the cognitive control hypothesis predicts that supervised automated driving, which mainly automates skill-based tasks, should not reduce workload for skilled drivers compared to manual driving.

2.1.2. MEASURING WORKLOAD

There is an extensive amount of literature reviewing methods to measure workload, e.g. (Cain, 2007; de Waard, 1996; Miller, 2001; Paxion et al., 2014; Stanton et al., 2005, Ch. 39; Stanton et al., 2013, Ch. 8; Young, Brookhuis, Wickens, & Hancock, 2015). Each measure is sensitive to a different set of resource pools, and in different performance regions (de Waard, 1996). Here we discuss measures used in the present study. The collection of workload measures can be classified into subjective rating (self-report) or objective measures (task-performance and physiological measures).

Subjective rating reflects workload as experienced by the operator (driver) and is thus sensitive to changes in effort. It is the simplest way to measure workload and is considered more reliable than physiological measures (Miller, 2001). The NASA task load index (TLX) (Hart, 2016) is a commonly used subjective measure in aviation and automotive research, and captures operator workload through six dimensions (mental, physical and, temporal demand; effort, frustration and performance) and reduces variability between participants and task contexts by letting participants score the relevance of each of these items. A variant, the Raw TLX (R-TLX), ignores this scoring step and has been found to remain an effective workload measure (Hart, 2016). We adopted the R-TLX to reduce the length of the post-drive questionnaire.

Subjective workload ratings have high face validity, but ratings may deviate from the actual workload. Stanton (1995) and Young and Stanton (1997) suggest in the contextual attention theory (CAT) that imbalance between perceived and actual demands and/or resources is one of the mechanisms through which poor performance can emerge, and that such an imbalance is especially likely in automated driving when there is insufficient feedback on the driver's performance (Norman, 1981). In order to capture such a mismatch, it is necessary to also collect objective measures of workload.

Objective workload measures often derive from task-performance, assuming reduced performance with under- and overload. Performance can either be measured on the primary task, or on secondary tasks. Most primary tasks in driving require manual operation of the vehicle (e.g. lane keeping performance), and are not suitable for automated driving. Secondary tasks aim to measure the driver's spare capacity. They tend to have a high reliability and can be designed to target specific resource pools. Consequentially, a variety of secondary tasks can be found throughout literature. One drawback of secondary tasks is that they interfere with the primary task. The detection-response task (DRT) is a secondary task designed to measure driver's cognitive load, and has been verified extensively (NEN-ISO 17488, 2016). Specifically, it measures the driver's ability to shift attention between the primary driving task and the DRT by measuring the delay between stimulus and response. When using a modality not interfering with the driving task (i.e. tactile or auditory), it is regarded as a pure measure of cognitive load. Compared to other secondary tasks, the additional cognitive demand induced by the DRT is generally considered to be low (Martens & van Winsum, 2000), but not effortless and not prone to automaticity (Engström et al., 2017). We selected the auditory DRT since monitoring of automation is centered in the cognitive resource pool, has low interference with the driving task and is not visually distracting. We preferred auditory over tactile stimuli, as this minimizes intrusive instrumentation.

Physiological measures sensitive to changes in workload include cardiovascular activity, galvanic skin response, brain activity and pupilometry. Brookhuis regards physiological measures as "*the most natural type of workload index, since, by definition, work demands physiological activity*" (cited in Stanton et al., 2005, p.17-2). Physiological measures can be recorded continuously and unlike performance measures they do not require any task to be performed, which makes them interesting for driver state monitoring. Cardiac monitoring is one of the most commonly used physiological measures of workload. Mental effort is associated with arousal which increases heart rate, while heart rate variability is found to decrease under high mental effort (Stanton et al., 2005, Ch.20, Ch.39). This relation between heart rate variability and mental effort is related to the sympathovagal balance between the sympathetic (0.02-0.06 Hz) and parasympathetic (0.15-0.40 Hz) nervous system, which is measured in the 0.10 Hz range, or as the ratio between high and low frequency ranges (though the idea that the LF/HF ratio is a suitable indicator for the sympatho-vagal balance has been challenged; see Billman (2013) for a comprehensive review). However, heart activity (and variability in particular) are not selective measures of workload. They primarily respond to the body's regulatory functions and are hypersensitive to noise from movement, changes in breathing rate and speech (Jorna, 1992; Young, 2000). We recorded heart activity and explored LF/HF ratio and standard deviation of inter-beat intervals, as they are related to mental workload and less affected by artifacts than other variability measures (Stapelberg, Neumann, Shum, McConnell, & Hamilton-Craig, 2017). Eye measures related to workload include blink rate, horizontal gaze dispersion (for highway driving) and pupil diameter (Marquart, Cabrall, & de Winter, 2015). The latter is particularly sensitive to high levels of cognitive load, but requires careful control of light conditions (Kahneman, 1973). In addition, eye tracking can provide further insight into the quality of monitoring (i.e. task performance) by assessing changes in glance frequency and durations to regions of interest (Kircher & Ahlstrom, 2017). We included eye tracking in our study to assess visual load and monitoring quality.

2.1.3. EMPIRICAL WORKLOAD IN AUTOMATED DRIVING

The theories of workload can help to explain and predict how automation and other moderators affect the driver's workload, but for quantitative effects we need to examine empirical findings. To this end, we selected studies addressing effects of driving automation, traffic complexity and automation experience.

The empirical review from de Winter, Happee, Martens, and Stanton (2014) summarizes workload findings from 32 studies comparing different levels of automation on the NASA TLX and RSME (Rating Scale Mental Effort). TLX responses were converted to a percentage scale for better comparison to RSME, with the lowest item mapped to 0% and the highest to 100% . Studies were mainly performed in simulators, and indicated a workload reduction of 21% on average from manual to automated driving. Six of the reviewed automated driving conditions could be considered SAE2 (Damböck, Weißgerber, Kienle, & Bengler, 2013; McDowell, Nunez, Hutchins, & Metcalfe, 2008; Saxby, Matthews, Warm, & Hitchcock, 2013; Schermers & Malone, 2014). With SAE2 automation workload was only 13.5% lower compared to manual driving. Ratings ranged from 23-66% for manual and from 23-40% for SAE2 automated driving.

The influence of traffic complexity on workload can be as large as the use of driving automation, with a 35% workload increase from low to high traffic complexity in manual driving (Teh, Jamson, Carsten, & Jamson, 2013). During supervised automated driving, traffic increases demands for the monitoring task (Jamson, Merat, Carsten, & Lai, 2013).

While task complexity increases demand, experience with automation may reduce it. Until recently, the influence of experience with automation could hardly be investigated due to the unavailability of automation-experienced drivers. Simulator studies on workload in automation often include a familiarization period, but the 15-30 minute exposure times are too short for the development of experience (Beggiato, Pereira, Petzoldt, & Krems, 2015). Some studies have approximated automated driving experience by using adaptive cruise control (ACC) experienced drivers (Larsson et al., 2014; Naujoks, Purucker, & Neukum, 2016) or developed special procedures to create experience through training (Krampell, 2016). Some effects of experience, such as the perceived risk and trust, may also be hard to study in simulators, which pose limitations on the perceptual fidelity (de Winter, van Leeuwen, & Happee, 2012; Hallvig et al., 2013). However, some recent studies measured mental workload during automated driving on the road.

Solís-Marcos, Ahlström, and Kircher (2018), measured visual secondary task performance in a Volvo S90 equipped with pilot assist (SAE2) and included both automation-inexperienced drivers and vehicle owners who had experienced the automation for 4.5 months on average before participation. In contrast to their expectations, they found that automation use increased the percentage of incorrect responses to the secondary task compared to manual driving, despite similar task completion rates in both conditions and longer glances towards the visual task with automation. TLX ratings of mental effort were high (79% in manual driving and 67% with automation use), which indicates that in supervised automation, secondary visual-motor tasks can be very demanding. Automation-experienced drivers gave shorter glances to the road compared to automation-inexperienced drivers in all conditions. They also gave longer glances at the secondary task, and this behavior was more pronounced during automated driving compared to manual driving, whereas the inexperienced drivers did not change glance time with automation use.

Banks and Stanton (2016) studied the workload of automation-inexperienced drivers during a short but engaging trip in a prototype supervised automated vehicle. In contrast to findings from simulators, the perceived workload was higher during automated driving (42%) compared to manual driving (27%). The participants' lack of prior training with the system, the additional tasks (performing three lateral maneuvers and answering an interview) and reported issues with the automation's behavior may all have contributed to the perceived workload increase.

Heikoop, de Winter, van Arem, and Stanton (2017) performed an on-road test with professional drivers familiar to supercars, but with no prior experience with lateral automation in a Tesla Model S on the highway, following a lead vehicle after 30 minutes of test-track training. A simple secondary task (counting bridges) was performed during part of the trip. The perceived workload during automated driving was rated very low overall (average of 19%), which is even below findings from simulator literature and reduced over time, suggesting that accustomization occurred during the trip. Accordingly, negative standardized change scores between the pre-drive and post-drive engagement ratings on the Dundee stress-state questionnaire suggest an overall disengagement during the drives.

Eriksson, Banks, and Stanton (2017) investigated the transition time in non-critical control transitions on the road in a Tesla Model S and compared it to a simulator study. Participants in the on-road experiment had prior experience with driving automation while the participants of the simulator study did not. Drivers in the on-road experiment regained control 32% (1.5 seconds) faster on average compared to the simulator drivers. The workload was perceived as low in both studies and no significant difference was found between the two studies.

Naujoks et al. (2016) performed a field study measuring secondary task uptake, secondary task workload and compensatory behavior in congested traffic while driving manually, with ACC and ACC plus steer assist in a Mercedes-Benz E-Class. They explored the effect of automation experience by comparing drivers with and without prior ACC experience. ACC-experienced drivers performed more secondary tasks in automated driving than in manual driving, in particular when driving at lower speeds, suggesting reduced workload with automation at lower driving speed. The effect however was not present for ACC-inexperienced drivers, suggesting that automation experience is a prerequisite for freeing cognitive resources for secondary tasks.

Based on these preceding works, we formulated the following hypotheses for supervised automation:

H1. Workload will be higher in the engaging condition than in the monotonic condition for both manual and automated driving.

H2. Automation will reduce workload.

H3. Workload during automated driving will be higher for automation-inexperienced drivers compared to automation-experienced drivers

We expect these effects to occur for both objective (auditory DRT) and subjective (R-TLX) workload measures. It should be noted that H2 concurs with a wide range of findings in various tasks, including driving in simulators, but not with the cognitive control hypothesis. Also, opposite effects were reported in two recent on-road studies as reviewed above.

2.2. METHODOLOGY

2.2.1. PARTICIPANTS

Two groups (N=8 each) of participants took part in the experiment and were selected through convenience sampling. Automation-experienced Tesla owners were recruited through the Dutch/Belgium section of the Tesla Motors forum (Tesla Motors, 2017). Seven reported using a Tesla and its Autopilot on a daily basis. One was an irregular user but reported 10,000 km travelled using Autopilot. One of the experts was the safety instructor, who had observed 8 participants prior to taking part himself.

The automation-inexperienced participants were invited through the universities' employee mailing list and through a list of drivers who had indicated their interest to participate in research regarding automated driving. Inexperienced drivers were required not to have experienced driving automation before. Users of adaptive cruise control were excluded but users of non-adaptive cruise control were included. The demographics of both groups are summarized in table 2.1.

Table 2.1: Demographics of the two participant groups, with mean μ, standard deviation σ and [interval].

	Experienced group	Inexperienced group
age	μ=43 σ=14 [27-69]	μ=41 σ=14 [21-61]
years licensed	μ=22 σ=15 [4-51]	μ=21 σ=15 [3-43]
km driven past 12 months	μ=26.500 σ=21.500 [7,500-75,000]	μ=15,000 σ=13,000 [3,000-42,500]
gender	7 male, 1 female	8 male

2.2.2. VEHICLE AND INSTRUMENTATION

An on-road driving task was performed with a rented Tesla model S 75D equipped with Autopilot (hardware version 1; update 8.0) and the driver's seat on the left side. The vehicle features supervised automation, which combines adaptive cruise control with automated lane keeping. The system supports lane changes (which have to be initiated by the driver) and adapts driving speed to traffic in the adjacent lanes and road curvature. The automation requires the driver to keep the eyes on the road and the hands on the wheel. An overview of the instrumentation can be seen in Figure 2.2. Video was recorded with three GoPro cameras observing the traffic in front and behind of the car, as well as the driver. A webcam observed the instrument panel.

a: Eye tracker
b: DRT button + heart activity sensor
c: Webcam
d: GPS antenna
e: Experimenters
f: IMU + DRT

Figure 2.2: Overview of the instrumentation

An auditory detection response task (DRT) was performed as an objective measure of the driver's cognitive workload. The DRT was implemented in Python on a Raspberry PI 3B running Raspbian Jessie. The implementation and analysis were in line with NEN-ISO 17488 (2016), with the following notable exceptions:

- An auditory stimulus was provided randomly with an on-set interval of 3-5 seconds with a 3.1 kHz tone lasting one second, irrespective of response time.
- Stimuli were presented over 5 minutes at a time (amounting to 72 stimuli per participant per condition).
- The button used to respond to the stimuli was strapped to the participant's right index finger, as the right hand had no driving-related tasks other than steering during the DRT.
- The DRT instruction was phrased as “Press the button as soon as you hear the signal, but keep your attention on the road".

Heart activity was recorded as a psychophysiological measure of arousal and workload. Two variability metrics were analyzed: standard deviation of inter-beat intervals (sdNN), where low variability indicates high workload; and low over high frequency ratio (LF/HF), where a high ratio indicates a high workload. These metrics were calculated every 30 s over 300 s of data.

Heart activity was recorded using an optical sensor mounted to the participant's right middle finger, powered by an Atmel AtMega328P embedded processor board. The sensor was able to obtain a heart rate measure, but occasionally suffered from artifacts (e.g. holding the steering wheel differently changed contact pressure of the sensor or reduced blood circulation in the fingertip). The heart rate and variability metrics were calculated using an open-source Python toolbox (van Gent, 2017; van Gent, Farah, van Nes, & van Arem, 2018). Data were collected at 100 Hz, and low pass filtered with a second-order Butterworth with a cutoff frequency of 5 Hz. The dominant (R-wave) peaks were identified as the maximum sample from any signal section rising above a 1.5 s moving average. Sections of poor data were identified by a variety of error detection and peak rejection algorithms, including the exclusion of heart rates outside the normal range [30-130 bpm] as well as any R-peaks whose associated inter-beat intervals exceeded the [250-300 ms] range.

A pupil labs head mounted eye tracker with the Linux distribution of pupil capture (pupil-labs v0.9.1, 2017) was included for the exploratory glance behavior analysis. However, a power outage on the second testing day resulted in software corruption, leading to random crashes of the tracking software. As a consequence, we only obtained full recordings of two participants and dropped the eye tracking from further analysis.

Vehicle motion (6 DOF acceleration, speed and location) was recorded using an MPU6050 IMU and GTPA013 GPS sensor connected to a second Atmel processor.

A safety instructor sat next to the participant and was proficient in the use of the Autopilot and experienced in introducing new drivers to the vehicle. During the drive, his tasks were to inform or warn the driver when needed, to help with the navigation and vehicle settings from the center console and to provide answers to technical questions. He was also allowed to engage in idle conversations except when instructions were given by the experimenter or during the DRT. The participant was allowed to initiate a conversation at any time. We did not inhibit speaking to maximize behavioral validity. By allowing participants the freedom to engage in conversation, the effects of experimenters' presence on behavior became more representative to having any other passenger.

To control for confounders that are inevitable in an on-road study, the DRT data was enriched by annotating events which may influence the response, such as lane changes, uninstructed (dis)use of the automation and verbal interactions. For each stimulus-response pair, the following classifications were made through manual annotation of the video footage:

- Lane change: ego vehicle undergoing a lane change or having indicators activated
- Use of Autopilot (on/off)
- Periods of congested traffic (vehicle or traffic speed slower than 75 km/h)
- Driver speaking (y/n)
- Other occupant speaking (y/n)

A stimulus was classified when these events occurred at any moment between the end of this stimulus and the end of the preceding stimulus.

To obtain an accurate record of the experienced traffic conditions, traffic flow (intensity) and traffic speed were logged from the NDW open data server (National Data Warehouse for Traffic Information, 2018) every minute. For each recording that contained both values, traffic density was calculated as lane-averaged intensity divided by lane-averaged traffic speed, where empty lanes were ignored. After pre-processing, the lane-average traffic data was interpolated to the GPS position and time, to obtain a continuous estimate of the traffic condition surrounding the vehicle. This interpolation accounted for differences in information travel in free and congested traffic as described in the traffic-adaptive model of Treiber and Helbing (2002). When GPS data was not available, a single average was computed for the road section over the duration of the condition.

2.2.3. SUBJECTIVE MEASURES

Three questions were asked while driving, to which the participant responded verbally on a scale from 1 to 9. The first question covered mental demand, the second regarded alertness and the third question reflected the driver's trust. In each condition, the three questions were asked before and after performing the DRT to verify that this task did not influence the driver's state. The questions were phrased as: *On a scale from 1 to 9, how mentally demanding was the* {*manual driving, use of autopilot*}*? On a scale from 1 to 9, how alert were you during the* {*manual driving, use of autopilot*}*? On a scale from 1 to 9, how much did you trust* {*yourself with the driving, the automation*}*?* For the alertness question, the descriptions of Karolinska sleepiness scale (KSS) (Kaida et al., 2006) were used: 1: very sleepy, great effort to keep awake 2: sleepy, some effort to keep awake 3: sleepy, but no effort to keep awake 4: some signs of sleepiness 5: neither alert nor sleepy 6: rather alert 7: alert 8: very alert 9: extremely alert. Each time, the participant was reminded of the description of the given response and was permitted to revise the response accordingly. The demand and trust questions were not anchored while driving, but 1 was described as low and 9 as high before departure.

The NASA Raw Task Load Index (R-TLX) was filled out after each driving condition on a 21-point scale. (we report results in percentages, with the lowest possible rating mapped to 0% and the highest possible rating mapped to 100%) Additionally, a confidence questionnaire based on Rendon-Velez et al. (2016) was used, with items (1) *driving manually was easy*, (2) *I felt confident to drive manually*, (3) *I had a feeling of risk*, (4) *using the automation was easy*, (5) *I felt confident to use the automation* and (6) *I had a feeling of risk during automated driving* on a 5-point scale with anchors: *disagree strongly, disagree a little, neither agree nor disagree, agree a little, agree strongly*. Also the 12-item automation trust questionnaire from Jian, Bisantz, and Drury (2000) was adopted on a 7-point scale. This questionnaire was only filled out for the drive as a whole, and not for each condition separately.

2.2.4. ENVIRONMENT

Two highway sections were selected to represent two levels of driving complexity; an engaging environment with a road geometry stimulating continuous traffic interaction, and a monotonic environment with lower traffic density and a simple road geometry and a low chance for high-demand scenarios to occur (Figure 2.3).

For the engaging environment, the A10 (ring-East of Amsterdam) was selected for its high traffic density throughout the day and the 10-13 on/off-ramps (depending on direction traveled). To maximize the traffic interaction, the driver was instructed to drive in the right lane as much as possible and was allowed to overtake slow moving traffic. On parts of this road it is legal to use the shoulder lane, but we instructed drivers to keep the regular right lane to avoid unpredictable behavior of the Autopilot. The A10 was entered from the A1 and followed down till exit Oud Zuid, either driving manually or using the automation. The highway was then followed in the opposite direction until exit Zeeburg, during which the DRT was performed. The route was then repeated with the remaining mode of automation.

For the monotonic environment the A6 between Almere (exit 7) and Lelystad (exit 10) was selected, which is a straight two-lane highway with low traffic density and no on/off ramps between the two cities. Here the driver was instructed to remain in the right most lane, to not overtake slow traffic and to drive as fast as traffic permits, but not faster than 110 km/h. Drivers got stuck behind a truck or trailer driving approximately 85 km/h in 80% of the monotonic scenarios. The automation was either used on the way towards Lelystad or back towards Almere.

The two driving environments were located 15 minutes away from one another. The A1 connects the two locations and was used for the familiarization. The A1 was entered from the A9 and first traveled in eastern direction. When the familiarization was to be followed by the engaging environment in Amsterdam, the first available exit was taken before practicing the automated lane change, but no later than Naarden. When the familiarization was to be followed by the monotonic condition, the road was simply continued towards the A6. The order of monotonic/engaging drives after the familiarization was counter balanced. For the inexperienced driver, the total trip lasted for 1.5 h when first driving to the monotonic condition or 1.75 h when first driving to the engaging condition. The experienced driver needed 1.5 h for either route due to the shorter familiarization.

The McDonald's Amsterdam Zuidoost was selected as the start/end point of the route, as it was logistically located between the highway entrance and the Tesla supercharger, and provided the facilities needed for welcoming the participants.

Figure 2.3: The driven route. Images were recorded during the drive of participant 5.

2.2.5. PROCEDURE

The experiment was approved by the human research ethics committee (HREC) of the TU Delft. Upon arrival, the participants were informed of the tasks and risks of the experiment. A pre-drive questionnaire was filled out and the procedures were explained. Prior to departure, the safety instructor informed the participants about the operation and limitations of the vehicle and the automation while the experimenter positioned the eye tracker, heart rate sensor, and DRT button. To explain and demonstrate the basic operation of the vehicle, the safety driver closely followed a checklist covering the controls; all possible automation modes, their functional meaning and methods for activating and disengaging them. Also the automation-related symbols were explained using Figure 2.4. Guidelines for safe use were phrased as: *when using autopilot, you should be on the lookout for things that the automation cannot handle correctly, for instance: lane markings that are not well visible or that have to be crossed, situations that require very strong braking or steering, traffic that behaves unexpectedly, when the car does something you would not do, or when it doesn't do something when you would.* The participants were further instructed repeatedly to remain attentive drivers at all times.

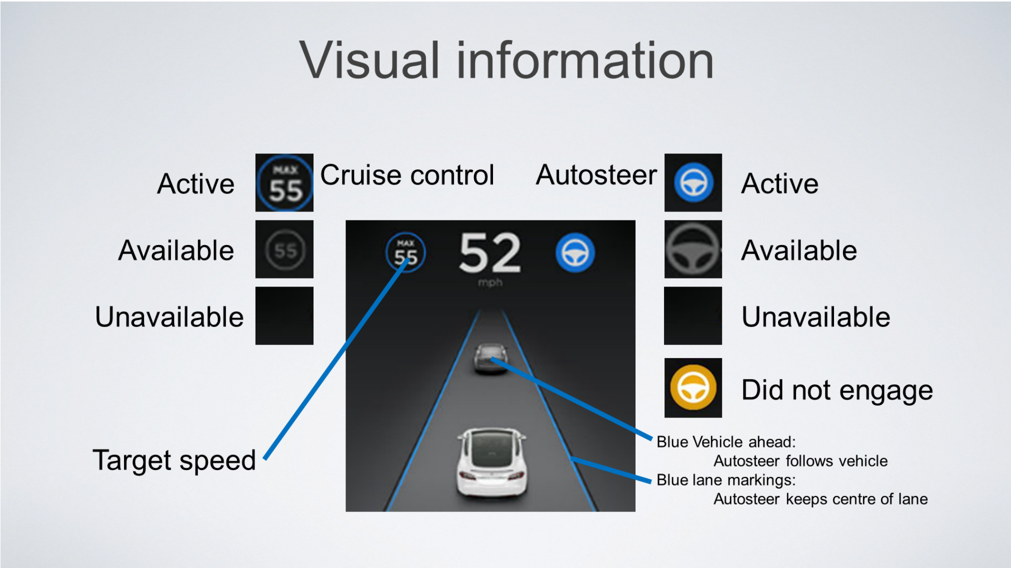


Figure 2.4: illustration of the automation-related information on the instrument panel.

Once on the highway, a familiarization drive was performed, during which the participants were introduced to the general operation of the vehicle and the basic behavior of the automation. The performed tasks covered the different methods of activation and deactivation of the automation, the adjustment of the cruise speed setting and the automated lane change. Questions were asked to make sure that the driver understood the instrument panel and operation of the vehicle. The familiarization lasted as long as necessary to let the participant perform each task successfully. The inexperienced drivers needed around 20 minutes while the experienced drivers required 8 minutes on average. The participants then drove manually to the engaging or monotonic environments, where they performed 4 rides in manual/automated and engaging/monotonic conditions in a randomized order.

To test for a possible effect of the DRT on subjective workload, within each condition we drove 5 minutes without DRT and 5 minutes with DRT. Each condition started with the instructions regarding driving behavior, followed by 5 minutes of driving without DRT. The three questions regarding mental demand, alertness and trust were asked before and after the DRT.

2.3. Analysis and Results

The drives were performed on workdays between the 3rd and 10th of March 2017. All drives took place between 9:00 and 16:45 to avoid rush hours and congestion. All tests occurred in normal (Dutch) weather conditions, except for two automation-experienced drivers, who drove in heavy rain. The automation operated reliably for all drivers in all experienced conditions. There was one occasion where ACC momentarily braked for no apparent reason (we suspect that the radar caught a guard rail at the end of a highway exit) and a couple of non-critical conflicts with surrounding traffic (e.g. another vehicle cutting in front, Autopilot attempting to undertake another vehicle at the right lane); all were adequately resolved by the participants with no or negligible inconvenience. Two experienced drivers did not follow all instructions during the monotonic condition. One occasionally overtook trucks. The other had not turned off ACC during manual driving in the first half of the monotonic drive.

During the DRT, the 8 automation-inexperienced users collectively made 8 lane changes in the engaging condition during the automated drive (2 with automation turned off) and 5 in the manual drive. The 6 experienced users made 17 in the automated drive (4 with automation turned off) and 15 in the manual drive, suggesting that the experienced users were more comfortable with making lane changes. In the automated engaging drive, the Autopilot was turned off 9.8% and 5.2% of the time by inexperienced and experienced drivers respectively, while in the monotonic drive the Autopilot was turned off 1.9% and 0.7% of the time respectively. The descriptives and effect sizes resulting from a 3-way ANOVA are provided in Table 2.4 and Table 2.1 and will be described in sections 2.3.2 and 2.3.3.

2.3.1. Traffic conditions

Despite a technical malfunction in the power supply to the GPS equipment, we managed to retain the GPS data of 48 5-minute conditions among 11 drives. Since the interpolated traffic estimates require this GPS signal, approximate traffic estimates were made for the conditions where the GPS signal was not available. The two methods were compared to ensure that the approximation is appropriate. Pearson's r and the difference in means between the two calculation methods as well as its significance according to a paired Student's t-test are outlined in Table 2.2. On all three metrics, the two calculation methods correlate well. Although the difference in means of traffic flow and (consequentially) density is statistically significant, we deemed its practical significance small enough to include the approximate traffic data in the analysis.

Table 2.3 shows the mean driving and traffic conditions among the experimental conditions, as well as the sample sizes for which GPS (and thus traffic speed) is available. Figure 2.5 shows

Table 2.2: Pearson's correlation between interpolated and approximated estimates of traffic speed, flow and density

	r	$E_{continuous\text{-}approx}$	t(1,47)	p
Traffic speed	0.836	-0.407 (km/h)	-0.415	.680
Traffic flow	0.871	-85.7 (veh./h)	-2.858	.006
Traffic density	0.929	-0.458 (veh./km)	-2.385	.021

traffic density and speed of the individual participants. It demonstrates that comparable traffic conditions were obtained between manual and automated driving for both groups, while differentiating between the two driving environments.

Table 2.3: Driving speed and traffic properties.

	Monotonic				Engaging			
	Automated		Manual		Automated		Manual	
	μ	σ	μ	σ	μ	σ	μ	σ
car speed [km/h]	87.3	13.7	88.5	10.5	87.2	16.6	90.0	9.9
traffic speed [km/h]	111.2	15.6	115.2	12.4	96	10.5	96.9	6.5
traffic flow [veh./h]	751	524	707	530	1303	421	1302	448
density [veh./km]	9.4	5.7	8.06	4.19	14.14	4.9	14.04	4.29
N_{used}	16		16		16		16	
N_{GPS}	8		8		8		4	
N_{GPS} during DRT	5		5		4		3	
$N_{traffic}$	16		16		14		13	
$N_{traffic}$ during DRT	14		14		13		11	

N_{used} = number of participants driving each condition.
N_{GPS} = number of participants for which GPS (and thus car speed) data is available.
$N_{traffic}$ = number of participants for which traffic data (speed, flow, density) is available.
During DRT = number of samples for which 5 minutes of data is available
while performing the DRT.

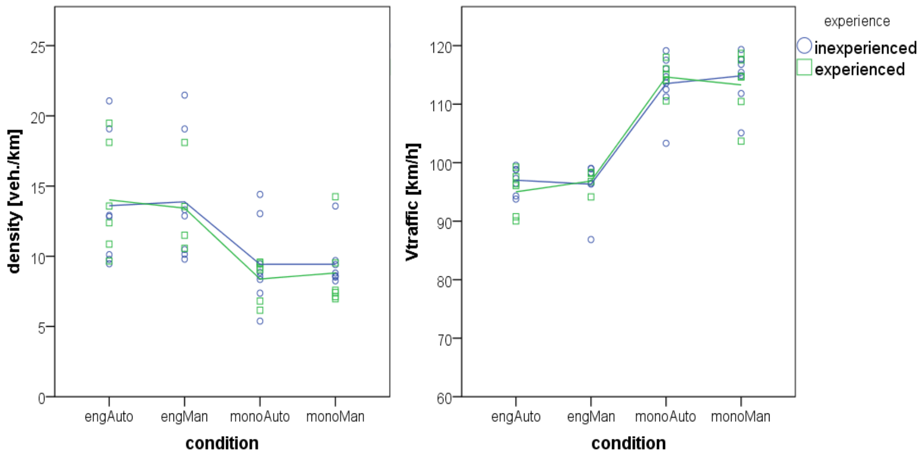


Figure 2.5: Traffic density (left) and speed (right) as experienced by the two groups across all test conditions during the DRT. Lines indicate condition means, symbols represent individual participants. engAuto = engaging environment using automation, engMan = engaging environment with manual driving; mono" = monotonic environment with either driving mode.

2.3.2. DETECTION RESPONSE TASK (DRT)

The auditory DRT was performed as an objective measure of changes in cognitive load. Due to missing values, the experienced and inexperienced group are represented by six and eight participants respectively.

Lane changes contributed to only 0.10% of the time for the inexperienced group and 0.17% of the time for the experienced group. During lane changes in the engaging condition reaction times were 240 ms ($t(5)=3.206$, $p=.024$) and 261 ms ($t(7)=3.797$, $p=.007$) slower for the experienced and inexperienced drivers respectively. These effect sizes are comparable to the 1-back task (adds 232 ms to baseline manual driving according to NEN-ISO 17488 (2016, Table E8)) or counting backwards from a 3-digit number (adds 125 ms to baseline manual driving according to Merat, Johansson, Chin, Nathan, and Victor (2006, figure 30)). Of all DRT misses, 30% and 25% occurred during lane changes for the inexperienced and experienced group respectively, resulting in a miss rate of 12% for automation-experienced and 19% for automation-inexperienced drivers during lane changes.

The time spent conversing varied across participants, with 3 out of 14 participants accounting for 65% of all conversations held. When speaking, the participants' reaction time was 240 ms slower compared to when being silent ($t(13)=3.45$, $p=.004$). No difference in reaction time was found between other occupants being silent or speaking ($t(13)=1.388$, $p=.188$). The DRT thus showed a higher miss rate and slower response time during lane changes and a slower response time during speaking. This indicates that single DRT responses can uncover additional information when combined with the identification of external events.

To remove some of the confounders for the DRT, we omitted all stimuli that occurred during lane changes, uninstructed automation (dis)use, and during driving or traffic speeds below

75 km/h. This refinement excluded 10.4% of the stimuli (with 17.0% of stimuli removed from the engaging-automated condition, 18.7% from the engaging-manual condition, 1.8% from the monotonic-automated condition and 0.5% from the monotonic-manual condition). Results after this removal are presented in Table 2.4 and Table 2.5. Stimuli during speaking or listening were not excluded from the analysis, as we consider speaking to be inherent to the driving strategy, but we verified if removing stimuli during speaking would change the trends and effects reported in Table 2.4 and Table 2.5. All trends with automation, experience and complexity remained. Reaction time received slightly smaller effect sizes, but significant effects remained significant. The effect of environment on miss rate became insignificant (p=.081). It is however not surprising that miss rates, which depend on the occurrence of rare events, are sensitive to data removal.

2

Reaction times and miss rates show similar trends (see Table 2.5). Both reaction time and miss rate have a main effect of environment (engaging environment has 99 ms slower reaction time and 1.6% higher miss rate compared to the monotonic environment), indicating that the engaging environment is more demanding than the monotonic environment. This effect size is similar to adding the 0-back task to baseline manual driving, which increases reaction time with 99 ms according to NEN-ISO 17488 (2016, Table E8). Reaction time also shows a main effect of automation (38 ms slower reaction during automated driving compared to manual driving), indicating that automation use resulted in less spare cognitive capacity compared to manual driving. Miss rate did not increase significantly with automation. However this can be attributed to a capping effect from the measurement resolution; with 72 stimuli, only miss rate increments of 1.4% can be distinguished with each measurement. No main effect for experience was found. None of the interactions was statistically significant.

Although the interaction automation$*$ environment is not statistically significant (p = .065), the increase in reaction time during automated driving is more pronounced in the engaging environment ($E_{\text{auto-man}}$ = 64 ms, SE = 18) compared to the monotonic environment ($E_{\text{auto-man}}$ = 13 ms, SE=16). Thus, automation seems to increase cognitive workload particularly in the engaging condition. These effect sizes are similar (but opposite) to comparing DRT performance with and without manual driving (baseline driving increases DRT reaction time with 52ms on average according to NEN-ISO 17488 (2016, Table E8)).

Table 2.4: DRT reaction time and miss rate, and R-TLX

		Reaction time (ms)				Miss rate (%)				R-TLX (%)			
		Experienced N=6		Inexperienced N=8		Experienced N=6		Inexperienced N=8		Experienced N=7		Inexperienced N=8	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Engaging	Automated	548	212	532	161	3.10	2.00	4.19	3.09	24.3	20.3	42.9	20.1
	Manual	491	232	460	160	2.55	2.15	3.54	2.99	48.3	19.6	42.6	17.1
Monotonic	Automated	379	109	451	182	1.70	0.62	1.93	0.70	10.2	7.1	24.7	16.2
	Manual	386	114	420	167	1.65	1.09	1.79	1.09	32.0	19.6	29.6	15.4

2.3.3. NASA R-TLX

Subjective workload ratings were collected using the NASA R-TLX. Due to missing values, the experienced and inexperienced group were represented by seven and eight participants

Table 2.5: Univariate effects and interactions of 3-way ANOVAs for reaction time, miss rate and overall workload. The effect of automation in the R-TLX contradicts that of RT and MR. Sub-items indicate which R-TLX sub-scale items were significant (<u>m</u>ental, <u>ph</u>ysical, <u>t</u>emporal, <u>pe</u>rformance, <u>e</u>ffort, <u>f</u>rustration)

	Reaction time (ms)			Miss rate (%)			R-TLX (%)			
Effects and interactions	F(1,12)	p	η_p^2	F(1,12)	p	ν_p^2	F(1,13)	p	η_p^2	Sub-items
environment	**15.168**	**.002**	**.558**	**7.036**	**.021**	**.370**	**14.584**	**.002**	**.529**	m,ph,t,pe,e
automation	**11.321**	**.006**	**.485**	0.480	.502	.038	**8.871**	**.011**	**.406**	m,ph,t,e
experience	0.028	.870	.002	0.972	.344	.075	0.886	.364	.064	-
environment*experience	2.266	.158	.159	0.514	.487	.041	0.003	.957	.000	-
automation*experience	1.306	.275	.098	0.007	.934	.001	**5.945**	**.030**	**.314**	m,pe,f
environment*automation	4.111	.065	.255	0.360	.559	.029	0.175	.682	.013	-
environment*automation* experience	0.218	.649	.018	0.000	.999	.000	1.119	.309	.079	-

respectively. The subjective workload is given in Table 2.4 and Figure 2.6, which also includes the six contributing items. The main effects and interactions are given in Table 2.5. For each effect and interaction, we indicated the sub-scale items that were statistically significant. The familiarization condition was excluded from further analysis, as it differs in road type.

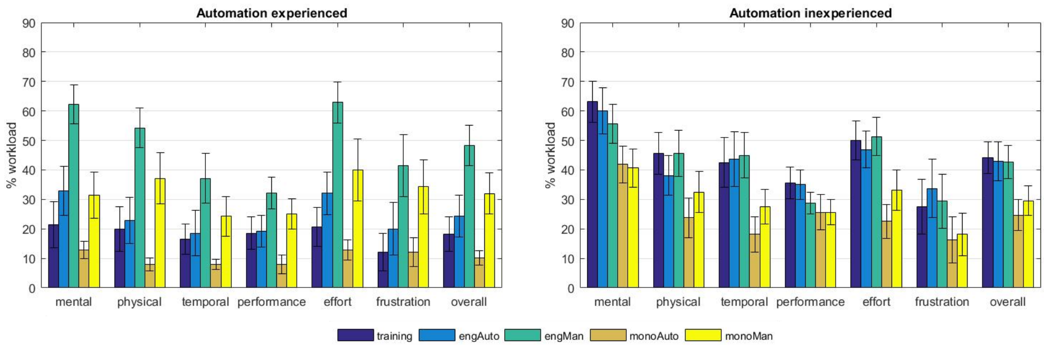


Figure 2.6: R-TLX for the experienced (left) and inexperienced (right) group, converted to percentage. Whiskers indicate standard errors.

Overall, workload was perceived 15.4% higher in the engaging than in the monotonic environment, and 12.6% lower during automated driving compared to manual driving. The interaction automation∗experience however indicates that only the experienced group ($E_{\text{man-auto}}$ = 22.9% , p=.003) perceived a workload reduction with automation use, whereas the inexperienced group did not experience a workload difference between manual and automated driving ($E_{\text{man-auto}}$ = 2.29% , p=.699). The perceived workload reduction with automation for the experienced group was consistent for both traffic conditions and for the six contributing items (Figure 2.6 left). No main effect of experience and no further interactions were observed.

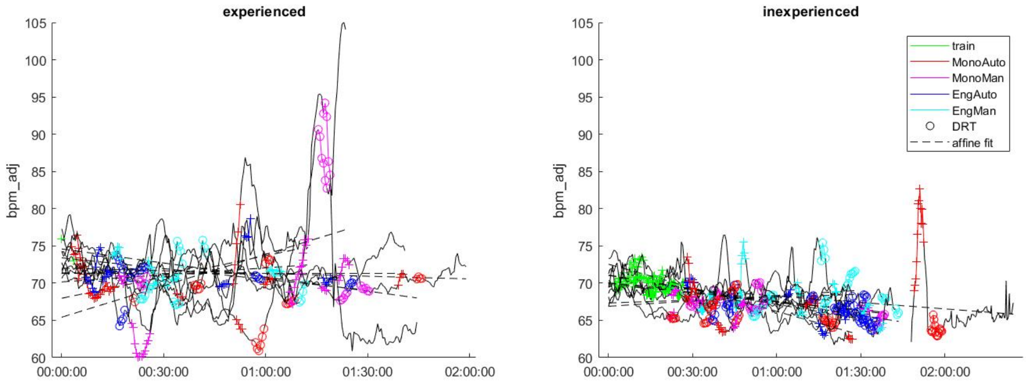


Figure 2.7: Mean-adjusted heart rate (bpm) over time of experienced (left) and inexperienced (right) drivers for each individual driver. Each sample is calculated over a 5 minute sliding window. Data collected during the conditions are highlighted.

2.3.4. HEART ACTIVITY

Individual heart rate traces are shown in Figure 2.7. An ANOVA on the mean-adjusted heart rate did not reveal any main effects or interactions on environment or automation use. Similarly, no effects or interactions were found for sdNN or LF/HF. Heart rate varied over time without apparent relation to the conditions or observed events. We suggest that the inherent variability of on-road driving along with artifacts from speaking, gripping of the steering wheel and other confounders overshadowed any possible effects resulting from automation or complexity of the environment.

A linear regression on the mean adjusted heart rate shows a time on task effect for the inexperienced group (b=-0.034 bpm/min, $F(1,169)=51.71$, $p<.001$), but not for the experienced group (b=-.002 bpm/min, $F(1, 138)=0.024$, $p=.878$). This indicates that the inexperienced drivers may have been acclimatizing to the vehicle and automation use during the experiment. Apart from this trend, heart rate proved ineffective to disclose significant effects of automation and traffic complexity.

2.3.5. QUESTIONNAIRES

To assess the impact of the DRT, the participants rated mental demand, sleepiness and trust both before and after performing the DRT. Overall, mental demand was 32.4% without the DRT and 36.9% with DRT, which is a measurable yet small increase in mental demand ($F(1,9)=3.361$; $p=.027$). No difference between driving with and without DRT was found for the KSS ($F(1,9)=0.941$; $p=.357$) or for trust ($F(1,9)=0.764$, $p=.405$).

Drivers reported lower sleepiness (KSS) in the engaging condition compared to the monotonic condition ($E_{mono\text{-}eng}=0.875$ points; $F(1,14)=18.08$, $p=.001$), but KSS was not affected by automation ($F(1,14)=1.577$, $p=.230$), experience ($F(1,14)=0.140$, $p=.714$) or on any of the interactions.

An overall confidence rating was computed over the items of the confidence questionnaire. Ratings ranged from 63% to 95% as can be seen in Table 2.6. The confidence questionnaire only showed a main effect of environment, with the engaging condition providing 13.4% less confidence than the monotonic condition (F(1,12)=13.38; p=.003). The ratings suggest that experienced drivers felt more confident during automated driving than in manual driving, while the inexperienced drivers felt more confident during manual driving than in automated driving, but the interaction automation∗experience was not statistically significant (F(1,12)=4.37; p=.059).

In the 12-item automation trust questionnaire, automation-experienced drivers reported a higher trust in the automation than the automation-inexperienced drivers. (experienced: μ=84.9% , σ=9.42; inexperienced: μ=67.7% , σ=14.0; t(1,13)=2.82; p=.014)

Table 2.6: Confidence ratings (%) among conditions and participant groups.

		Experienced N=7		Inexperienced N=7	
		μ	σ	μ	σ
Engaging	Automated	91.9	11.7	63.0	30.4
	Manual	67.9	20.0	73.7	28.4
Monotonic	Automated	95.2	12.5	80.9	19.8
	Manual	84.6	24.2	89.3	14.1

2.3.6. POST-HOC ANALYSIS

To complement the ANOVAs presented in table 2.5, we checked post-hoc for any relations between traffic conditions, heart rate (BPM, sdNN, LF/HF) and DRT response time within the experimental conditions. Pearson's correlation was used to explore relations between the participant averages in each condition. Traffic density correlated inversely with traffic flow on both environments (Pearson's r=-.590 for engaging; r=-.624 for monotonic), which is in agreement with traffic flow models. Within the conditions, none of the traffic metrics correlated with neither RT nor with any of the heart rate metrics.

We further checked if the difference in mileage between participant groups could confound the results by observing the correlations between the DRT and R-TLX measures for both groups and all conditions. As shown in Table 2.7, 19 out of 24 correlations were of negative sign, which suggests a higher mileage is associated with better performance at the secondary task. However, the sample sizes are too small to make any conclusive statements regarding the effect of mileage. Although the miss rate correlation of the experienced group in the monotonic condition during automation was statistically significant (p=.016), this may be attributed to an inflated type I error from the multiple comparisons being made. A Bonferroni correction for the 8 comparisons would dictate that the probability is to be tested at a confidence of $\alpha' =$ 0.0064 instead of $\alpha = 0.05$, in which case also this correlation is not statistically significant.

Because the dynamics of mental demand vary at a smaller time scale than a five-minute average can reveal, we further explored relations among the measurements at a shorter time scale by looking for correlations between individual DRT responses, heart activity and traffic condi-

Table 2.7: Correlations between mileage and reaction time, heart rate and R-TLX. engAuto = engaging environment using automation, engMan = engaging environment with manual driving; mono" = monotonic environment with either driving mode. The *p*-values are not corrected for the 8 independent comparisons.

		Reaction time		Miss rate		R-TLX	
		Pearson's r	*p*	Spearman's ρ	*p*	Pearson's r	*p*
Inexperienced	engAuto	−0.153	.717	−0.621	.100	−0.348	.399
	engMan	0.039	.927	−0.185	.660	−0.451	.262
	monoAuto	0.233	.578	−0.113	.791	−0.241	.565
	monoMan	0.202	.631	0.296	.476	−0.403	.323
Experienced	engAuto	−0.577	.231	0.174	.742	−0.351	.440
	engMan	−0.441	.382	−0.696	.125	−0.078	.868
	monoAuto	−0.292	.574	−0.893	.016	−0.323	.480
	monoMan	−0.395	.439	−0.585	.222	0.583	.170

tions, as well as time on task effects. Since the raw response times are not normally distributed, Spearman's non-parametric correlation was used. No correlations were found between heart rate and the individual responses. Similarly, relations between the traffic conditions and individual responses did not reveal further relations within the driving conditions. No time on task effects were found for DRT reaction time using linear regression (b=0.1ms/stimulus, *p*=.645).

2.4. Discussion

In this study, we investigated how workload changes with attentively monitored automated driving in real-world conditions, and how this change is moderated by traffic complexity and by the driver's prior experience with automated driving.

The engaging traffic environment resulted both in a higher overall subjective (R-TLX) and objective (DRT reaction time and miss rate) workload compared to the monotonic environment. Additionally, the drivers remained as sensitive to changes in driving complexity while using automation as they were while driving manually. This supports hypothesis H1 and demonstrates that monitoring automation imposes a considerable task demand.

Hypothesis 2, reduced workload with automation, is only supported for the perceived overall workload (R-TLX) for automation-experienced drivers but not for automation-inexperienced drivers. Furthermore in both driver groups the objective cognitive load (DRT) increased with automation. These results were unexpected and show opposing effects on subjective and objective workload in the experienced drivers.

The R-TLX ratings suggest that automation experience is a prerequisite for a reduction in perceived workload. The automation-experienced user may have developed a less demanding (or automated) strategy for monitoring the automation, while the inexperienced driver may stay closer to strategies from manual driving. This view is also supported by Solís-Marcos et al. (2018), who showed that automation-inexperienced and -experienced drivers have different glance behavior and that only the automation-experienced group changes glance behavior with automation use. In our study, during automated driving, automation-experienced drivers did not perform better on the DRT task compared to automation-inexperienced drivers. Although we believe our per-group sample size is rather small to formally test hypothesis H3, we would like to point out that this observation also aligns with Solís-Marcos et al. (2018), who despite the longer glances of experienced drivers to the secondary task did not find a difference in task performance between the two groups. In contrast, Naujoks et al. (2016) found that automation use resulted in a higher secondary task completion rate compared to manual driving for ACC experienced drivers. A possible explanation could be the difference in driving speed, since higher performance in Naujoks et al. (2016) only occurred in slow-moving (<60 km/h) congested traffic.

It should be emphasized that our findings for automation-inexperienced drivers are in conflict with results from simulator studies as reviewed in de Winter et al., (2014), where automation-inexperienced drivers reported workload reduction due to automation of a magnitude similar to the reduction we found for experienced drivers. Since the automation-inexperienced group reported lower trust and confidence in the automation compared to the automation-experienced group, we suspect that this difference between real-world and simulator findings relates to the low validity of risk perception in driving simulators. This further suggests that driver trust accounts for a large difference in perceived workload reduction by automation. We should however remark that on-road studies provide mixed results. Automation-inexperienced drivers perceived low workload during automation in Heikoop et al. (2017), while Solís-Marcos et al. (2018) found high workload ratings for both inexperienced and experienced users of automation. With the emerging on-road studies addressing workload in automated driving with automation-experienced users, we believe that a new meta-analysis on the effects of automated driving may be in order.

The dissociation between perceived overall workload (R-TLX) and objective cognitive load (DRT) for the effects of automation deserves further examination. Although overall workload incorporates more than cognitive load, we believe that a direct comparison between the two measures is fair, because the monitoring sub-task, which is centered in the cognitive resource pool, forms a large part of the driving task, and because the mental demand sub-items of the R-TLX show similar patterns as the overall load. The increase in reaction time and miss rates could theoretically be attributed to mental underload since such performance reduction is an indicator of vigilance decrement (Greenlee et al., 2018). We have however several indications that this is not the case. The Karolinska sleepiness scale did not indicate any development of drowsiness, and when drowsiness had been compensated with state-related effort, we should have seen this reflected in the effort or mental demand sub-scales of the R-TLX. Furthermore, the periods of automated driving were relatively short (10 minutes of automated driving at a time, interluded with verbal ratings after 5 minutes). Although vigilance decrement can develop in such time span, we should have been able to see such decrement as a time-on-task effect, which we did not in our regression analysis. Finally, the reaction time increase with automation use was larger in the engaging condition compared to the monotonic condition for both driver groups, which contradicts the hypothesis that longer reaction times of this study signify underload.

The increase in DRT reaction time can also not be fully explained by the malleable attentional resource theory (Young & Stanton, 2002; 2007), which suggests that total capacity reduces when task demands are low. In order to explain the *reduction* in DRT performance, the capacity reduction should have been larger than the reduction in primary task demand, whereas Young and Stanton (2002) propose that spare mental capacity should still improve, but disproportionally to the reduction in primary task demand. The cognitive control hypothesis cannot explain the increase in reaction time, but provides an explanation why supervised automation did not reduce objective workload.

The increase in objective cognitive load (DRT) suggests that the ratio between task demand and allocated resource increases with automation, whereas the reduction in subjective workload suggests that this increase is not perceived as such. Assuming that the TLX ratings are not confounded by confirmation bias or attribution error, we believe that this is caused by a mismatch between perceived and actual workload as suggested in contextual attention theory. Stanton (1995) proposed there can be a mismatch between the perceived and actual demands, between perceived and actual resources, or between the perceived demands and perceived resources. Under-estimation of task demand should result in too few resources being mobilized for both primary and secondary task. Such underestimation would reduce monitoring performance but may still lead to improved secondary task performance, unless the primary task demand is met with a resource allocation that is higher than the perceived requirement. An over-estimation of allocated resource is particularly likely in low-effort conditions, and could explain the reduced DRT performance under lower perceived workload. We can further expect workload to be rated lower than actual load, when a surplus demand investment (either perceived, or allocated in response to the instruction to monitor attentively) is ignored or weighted less in the overall workload rating. This suggests that either 1) automation increases demand while automation-experienced drivers perceive less effort and lower demands, or that 2) automation does not reduce demand as much as we think, and we allocate a larger fraction of our resource to monitoring than we made available for it. The first suggestion would be in conflict with hypothesis H2 whereas the second is not. Both however indicate a difference between perceived and actual workload. While drivers remain sensitive to changes in task demand (i.e. changes in traffic), they appear to over-estimate their resource allocation. The idea that experienced users under-estimate the actual task load has implications for safe usage of SAE2 automation. It indicates that supervised automation does not increase spare capacity as much for secondary tasks or interaction with in-vehicle information systems as the driver believes. These perceptual differences should be incorporated in the design and user-education for these systems. The results also reinforce the importance of measuring workload both objectively and subjectively. We recommend to incorporate objective measures of both primary task performance (i.e. monitoring) and spare capacity when studying mismatches between perceived and actual workload.

Our DRT findings suggest that attentively supervised automation results in a healthy workload (i.e. a little higher than manual driving) and thereby do not support the concern that supervised automated driving causes mental underload. In contrast, in particular the automation-experienced drivers perceived a reduced workload with automation. This means that, when drivers supervise attentively, they can maintain a healthy workload while perceiving a meaningful comfort benefit. However the mismatch between objective and perceived load may be a point of concern when it motivates users to pay less attention than is required, which in turn could mediate underload and may compromise safety.

Limitations: Eye tracking could not be assessed due to technical malfunction. The number of participants was limited, which in particular makes the between-subject comparisons sensitive to individual differences. However, the within-subject effects were very consistent among participants, and persisted when correcting for speaking during the experiment. The instructions and presence of a safety instructor and experimenters motivated attentive supervision of the automation. The results should therefore be regarded as workload under intended use, which may differ from every-day use. The idle conversations held may have reduced the sensation of being in an experiment, raised energetic state and increased workload. Although these aspects are representative for a drive with other occupants, results may differ when driving alone, with no one to talk to. The effects of supervised automation with longer periods of automated driving in a naturalistic setting without additional motivations to supervise remain to be investigated. Convenience sampling balanced years licensed, age and sex. Mileage was not balanced between groups, but sample size was too low to correlate this between-group difference. Furthermore, the Tesla users were sampled from a forum which actively discusses the limitations and abilities of the vehicle. The owner's disposition towards the vehicle may have resulted in confirmation bias or attribution error (i.e. general satisfaction being expressed on the workload scale). As an approach to compensate for such rater bias in future experiments, we propose to assign automation users to different vehicle brands, or to group participants based on their disposition regarding the automation.

Highlights

- Monitoring automation imposes a healthy workload.
- This demand increases with driving complexity.
- Automation lowered subjective workload, but experience was a prerequisite for this.
- Automation lowered subjective workload, but increased objective cognitive load.
- This suggests that drivers underestimate the resources allocated.
- Studies should assess workload subjectively as well as objectively.

Acknowledgment

We kindly thank Paul van Gent for providing us with the heart rate and GPS equipment, and his support in pre-processing the cardiovascular recordings.

Funding: This work was supported by the NWO-TTW Foundation, the Netherlands, under the project "From Individual Automated Vehicles to Cooperative Traffic Management - Predicting the benefits of automated driving through on-road human behavior assessment and traffic flow models (IAVTRM)" -STW#13712.

2.5. REFERENCES

Banks, V. A., & Stanton, N. A. (2016). Keep the driver in control: Automating automobiles of the future. *Applied Ergonomics, 53 Pt B, pp. 389–395. https://doi.org/10.1016/j.apergo.2015.06.020*

Beggiato, M., Pereira, M., Petzoldt, T., & Krems, J. (2015). Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation*

Research Part F: Traffic Psychology and Behaviour, 35, pp. 75–84. https://doi.org/10.1016/j.trf.2015.10.005

Billman, G. E. (2013). The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. *Frontiers in Physiology, 4(26), pp. 1–5. https://doi.org/10.3389/fphys.2013.00026*

Cain, B. (2007). A review of the mental workload literature. *Defence Research and Development Canada Toronto Human System Integration Section, pp. 1–35.*

Damböck, D., Weißgerber, T., Kienle, M., & Bengler, K. (2013). Requirements for cooperative vehicle guidance. *IEEE Annual Conference on Intelligent Transportation Systems (ITSC), 16, pp. 1656–1661.*

De Waard, D. (1996). The measurement of drivers' mental workload. Traffic Research Centre VSC.

De Winter, J. C. F., Happee, R., Martens, M. H., & Stanton, N. A. (2014). Effects of adaptive cruise control and highly automated driving on workload and situation awareness, a review of the empirical evidence. *Transportation Research Part F: Traffic Psychology and Behaviour, 27(B), pp. 196–217.*

De Winter, J. C. F., van Leeuwen, P. M., & Happee, R. (2012). Advantages and disadvantages of driving simulators: a discussion. *Proceedings of Measuring Behavior, pp. 47–50.*

Engström, J., Markkula, G., Victor, T., & Merat, N. (2017). Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human Factors, 59(5), pp. 734–764. https://doi.org/10.1177/0018720817690639*

Eriksson, A., Banks, V. A., & Stanton, N. A. (2017). Transition to manual: Comparing simulator with on-road control transitions. *Accident; Analysis and Prevention, 102, pp. 227–234. https://doi.org/10.1016/j.aap.2017.03.011*

Greenlee, E. T., DeLucia, P. R., & Newton, D. C. (2018). Driver vigilance in automated vehicles: Hazard detection failures are a matter of time. *Human Factors and Ergonomics Society, 18720818761711. https://doi.org/10.1177/0018720818761711*

Hallvig, D., Anund, A., Fors, C., Kecklund, G., Karlsson, J. G., Wahde, M., & Akerstedt, T. (2013). Sleepy driving on the real road and in the simulator–A comparison. *Accident; Analysis and Prevention, 50, pp. 44–50. https://doi.org/10.1016/j.aap.2012.09.033*

Hart, S. G. (2016). Nasa-Task Load Index (NASA-TLX); 20 years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50(9), pp. 904–908. https://doi.org/10.1177/154193120605000909*

Heikoop, D. D., de Winter, J. C. F., van Arem, B., & Stanton, N. A. (2019). Acclimatizing to automation: driver workload and stress during automated car following in real traffic: *Transportation research part F: traffic psychology and behaviour, 65, pp. 503-517. https://doi.org/10.1016/j.trf.2019.07.024.*

Hirose, T., Kitabayashi, D., & Kubota, H. (2015). Driving characteristics of drivers in a state of low alertness when an autonomous system changes from autonomous driving to manual driving. *SAE Technical Paper 2015-01-1407. Advance online publication. https://doi.org/10.4271/2015-01-1407*

Jamson, A. H., Merat, N., Carsten, O. M.J., & Lai, F. C.H. (2013). Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions. *Transportation*

2

Research Part C: Emerging Technologies, 30, pp. 116–125. https://doi.org/10.1016/j.trc.2013.02.008

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4(1), pp. 53–71. https://doi.org/10.1207/S15327566IJCE0401_04*

Jorna, P.G.A.M. (1992). Spectral analysis of heart rate and psychological state: a review of its validity as a workload index. *Biological Psychology. (34), pp. 237–257.*

Kahneman, D. (1973). *Attention and effort. New Jersey: Englewood cliffs.*

Kaida, K., Takahashi, M., Akerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., & Fukasawa, K. (2006). Validation of the Karolinska sleepiness scale against performance and EEG variables. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology, 117(7), pp. 1574–1581. https://doi.org/10.1016/j.clinph.2006.03.011*

Kircher, K., & Ahlstrom, C. (2017). Evaluation of methods for the assessment of attention while driving. *Accident; Analysis and Prevention. Advance online publication. https://doi.org/10.1016/j.aap.2017.03.013*

Krampell, M. A. H. (2016). Accelerated behavioural adaptation through targeted training programs – the case of highly automated driving (Master's thesis). Department of Computer and Information Science (IDA) at Linköping University.

Kyriakidis, M., de Winter, J. C. F., Stanton, N., Bellet, T., van Arem, B., Brookhuis, K., Martens, M., Bengler, K., Andersson, J., Merat, N., Reed, N., Flament, M., Hagenzieker, M., Happee, R. (2017). A human factors perspective on automated driving. *Theoretical Issues in Ergonomics Science, 53, pp. 1–27. https://doi.org/10.1080/1463922X.2017.1293187*

Larsson, A. F.L., Kircher, K., & Andersson Hultgren, J. (2014). Learning from experience: Familiarity with ACC and responding to a cut-in situation in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour, 27 https://doi.org/10.1016/j.trf.2014.05.008*

Marquart, G., Cabrall, C., & de Winter, J. C. F. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing, 3, pp. 2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783*

Martens, M., & van Winsum, W. (2000). Measuring distraction: the peripheral detection task. *TNO Human Factors.*

McDowell, K., Nunez, P., Hutchins, S., & Metcalfe, J. S. (2008). Secure mobility and the autonomous driver. *IEEE Transactions on Robotics, 24(3), pp. 688–697. https://doi.org/10.1109/TRO.2008.924261*

Merat, N., Johansson, E., Chin, E., Nathan, F., & Victor, T. (2006). Specification of a secondary task to be used in safety assessment of IVIS.

Miller, S. (2001). Literature review: Workload measures. *National Advanced Driving Simulator. Iowa City, United States.*

National Data Warehouse for Traffic Information. (2018). NDW open actuele dataservice. Retrieved from http://opendata.ndw.nu/

Naujoks, F., Purucker, C., & Neukum, A. (2016). Secondary task engagement and vehicle automation – Comparing the effects of different automation levels in an on-road experiment. *Transportation Research Part F: Traffic Psychology and Behaviour, 38, pp. 67–82. https:*

//doi.org/10.1016/j.trf.2016.01.011

NEN-ISO 17488. (2016). ISO 17488:2016: DRT for assessing attentional effects of cognitive load. *Nederlands Normalisatie Instituut.*

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88(1).*

Patten, C. J. D., Kircher, A., Ostlund, J., Nilsson, L., & Svenson, O. (2006). Driver experience and cognitive workload in different traffic environments. *Accident; Analysis and Prevention, 38(5), pp. 887–894. https://doi.org/10.1016/j.aap.2006.02.014*

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in Psychology, 5, 1344. https://doi.org/10.3389/fpsyg.2014.01344*

Pupil-labs v0.9.1. (2017). Pupil Capture, Player, and Service release v0.9.1. Retrieved from https://github.com/pupil-labs/pupil/releases/tag/v0.9.1

Rendon-Velez, E., van Leeuwen, P. M, Happee, R., Horváth, I., van der Vegte, W. F., & de Winter, J. C. F. (2016). The effects of time pressure on driver performance and physiological activity: A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour, 41(41), pp. 150–169. https://doi.org/10.1016/j.trf.2016.06.013*

SAE International (2016). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. (J3016 SEP2016).*

Saxby, D. J., Matthews, G., Warm, J. S., & Hitchcock, E. M. (2013). Active and passive fatigue in simulated driving: Discriminateing styles of workload regulation and their safety impacts. *Journal of Experimental Psychology: Applied.*

Schermers, G., & Malone, K. M. (2014). *Dutch evaluation of Chauffeur Assistant (DECA): Traffic flow effects of implementation in the heavy goods vehicles sector.*

Solís-Marcos, I., Ahlström, C., & Kircher, K. (2018). Performance of an additional task during level 2 automated driving: An on-road study comparing drivers with and without experience with partial automation. *Human Factors: the Journal of the Human Factors and Ergonomics Society, 001872081877363. https://doi.org/10.1177/0018720818773636*

Stanton, N. A. (1995). Ecological ergonomics: Understanding human action in context. *Contemporary Ergonomics, pp. 62–67.*

Stanton, N. A., Hedge, A., Brookhuis, K., Salas, E., & Hendrik, H. (2005). *Handbook of human factors and ergonomics methods. Boca Raton: CRC Press.*

Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., & Jenkins, D. P. (2013). *Human factors methods: practical guide for engineering and design (2nd edition). Wey Court East, Union Road, Farnham, Surrey, GU9 7PT England: Ashgate Publishing Limited.*

Stapelberg, N. J. C., Neumann, D. L., Shum, D. H. K., McConnell, H., & Hamilton-Craig, I. (2017). The sensitivity of 38 heart rate variability measures to the addition of artifact in human and artificial 24-hr cardiac recordings. *Annals of Noninvasive Electrocardiology : the Official Journal of the International Society for Holter and Noninvasive Electrocardiology, Inc, 23(1). https://doi.org/10.1111/anec.12483*

Teh, E., Jamson, S., Carsten, O., & Jamson, H. (2013). Temporal fluctuations in driving demand: The effect of traffic complexity on subjective measures of workload and driving performance. *Transportation Research Part F: Traffic Psychology and Behaviour, 22, pp. 207–217. https://doi.org/10.1016/j.trf.2013.12.005*

2

Tesla Motors. (2017). Tesla Motors Club Forum: Belgium and the Netherlands. Retrieved from https://teslamotorsclub.com/tmc/forums/belgium-and-the-netherlands.118/

Treiber, M., & Helbing, D. (2002). Reconstructing the spatio-temporal traffic dynamics from stationary detector data. *Cooperative Transportation Dynamics, 3.1-3.24.*

Van Gent, P. (2017). Python heart rate analysis toolkit. Retrieved from https://github.com/paulvangentcom/heartrate_analysis_python

Van Gent, P., Farah, H., van Nes, N., & van Arem, B. (2018). Analysing noisy driver physiology real-time using off-the-shelf sensors: heart rate analysis software from the taking the fast lane project. *(Submitted for Publication to the Journal of Open Research Software). Advance online publication. https://doi.org/10.13140/RG.2.2.24895.56485*

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors, 50(3), pp. 433–441. https://doi.org/10.1518/001872008X312152*

Wickens, C. D. (1981). Processing resources in attention, dual task performance and workload assessment.

Wright, T. J., Samuel, S., Borowsky, A., Zilberstein, S., & Fisher, D. L. (2016). Experienced drivers are quicker to achieve situation awareness than inexperienced drivers in situations of transfer of control within a Level 3 autonomous environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60(1), pp. 270–273. https://doi.org/10.1177/1541931213601062*

Young, M. S., & Stanton, N. A. (2007). What's skill got to do with it? Vehicle automation and driver mental workload. *Ergonomics, 50(8), pp. 1324–1339. https://doi.org/10.1080/00140130701318855*

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics, 58(1), pp. 1–17. https://doi.org/10.1080/00140139.2014.956151*

Young, M. S., & Stanton, N. A. (1997). Automotive automation: Investigating the impact on driver mental workload. *International Journal of Cognitive Ergonomics, 1(4), pp. 325–336.*

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. *Human Factors, 44(3), pp. 365–375. https://doi.org/10.1518/0018720024497709*

Young, M. S. (2000). Attention, automaticity, and automation: Perspectives on mental underload and performance.

3

Exploration of the impact of SAE2 automation on driving behaviour: A naturalistic driving study

Jork Stapel, Riender Happee, Michiel Christoph, Nicole van Nes, Marieke Martens

This chapter represents preliminary results using only a part of the collected data. Finishing the analysis is currently not possible due to technical problems of the data-holder and Covid-19.

Abstract

To better understand the safety impact of automated driving, situational and longitudinal adaptation of automation use and driver attention have to be analysed in naturalistic settings. This study reports preliminary findings on automation use and driver attention from a longitudinal naturalistic driving study, for which data enrichment through visual annotation was automated and validated. The dataset is unique in that it includes one month of manual driving followed by two months with automation use, allowing for a longitudinal within-subjects analysis of behavioural adaptation. From among five vehicle types used in the dataset, this study examined Tesla and BMW production vehicles with adaptive cruise control (ACC) and lane keeping (LK) features, and compares one month of baseline driving to the first two months of automated driving.

Data enrichment successfully retrieved automation status from video for Tesla vehicles with an accuracy of 99.3% while automation of the BMW could be retrieved from CAN data. Head pose was obtained from low end cameras to automatically classify visually attended regions. While performing on-par with literature, head pose without gaze information was found to be insufficient for attention classification and head pose variance (horizontal & vertical) was selected as alternative measure for monitoring activity.

On the highway, ACC+LK was used 63% and 70% of the time for respectively the Tesla and BMW and occurred the least (49%) when driving 60-70 km/h. Both ACC and ACC+LK are used significantly more often with higher speeds while on highways. On roads with speed limits below 70 km/h, automation was used less than 8%, and use on urban roads was incidental rather than habitual. Usage did not change with time in trip, time of day or experience. In the experimental phase, head heading variance was smallest during ACC use, but did not differ between ACC+LK and baseline manual driving. Head pitch deviation increased over the first 6 weeks of automation use for ACC+LK use and head heading tended to increase during ACC and ACC+LK, which hints at behavioural adaptation.

Because part of the required data was either unavailable or inaccessible at the time of analysis, the preliminary findings reported here are limited to 6 participants for the descriptives and 3 participants for the statistics on Tesla and BMW vehicles. This manuscript will be extended including further participants and vehicles when this data becomes accessible.

3.1. INTRODUCTION

Supervised driving, or Level 2 automation (SAE International, 2018) is rapidly deployed in commercial cars. Level 2 systems can take over the continuous lateral and longitudinal control of the vehicle, but their limitations require the driver to actively monitor the driving task to ensure traffic safety. While these systems are intended to improve safety and comfort, drivers may not be fully aware of their vehicle's capabilities (Harms et al., 2020) and may also introduce new risks, such as driver inattention and incorrect expectations about the driver responsibilities, possibly leading to under-use or unsafe use. This has resulted in several accidents (Dijsselbloem et al., 2019). While these confirm that drivers are not always monitoring the environment sufficiently to intervene in time when a system fails, it remains unknown under which conditions and how attentively drivers normally use automation. Safety benefits of adaptive cruise control (ACC) and automated lane keeping (LK) as currently deployed in Level 2 automation remain unknown (Dijsselbloem et al., 2019; de Winter et al., 2014; Vlakveld, 2019). The evaluation is further challenged since system functionality, capability and interfaces differ strongly between brands and changes with over the air updates.

While Level 2 automation is active, the driver has to supervise the automation, and intervene when needed to ensure safety. Safe use requires that the driver is aware of these responsibilities, has an accurate understanding on how the vehicle may respond to the situation at hand, and maintains sufficient situation awareness to respond when necessary.

If and how a driver experiences these requirements can be inferred from how often and in which situations these systems are used, and from how drivers distribute visual attention between driving-related and other tasks. Several studies have used these measures to evaluate the safety of automated driving features. Jamson et al. (2013) found that the use of driving automation changed the driving style compared to manual driving and reduced the number of lane changes. Drivers spent more time on secondary tasks but adjusted their attention to the road depending on traffic. Similarly, Naujoks et al. (2016) demonstrated in a 2013 Mercedes E-class that only drivers with prior ACC experience perform more secondary tasks while using driving automation. Farah et al. (2020) found that drivers over-estimated the operational design domain as defined by the vehicle manufacturer during an on-road study with a Tesla model S. Banks et al. (2018) performed thematic video analysis of behaviours observed during on-road driving in a Tesla model S and identified multiple occurrences of missed notifications from the HMI leading to mode confusion.

The distribution of visual attention between driving-related and secondary tasks can be inferred from gaze or head movement (Lee et al., 2018) and provides guidelines for the risk of driver distractions (Strickland, 2013). Park et al. (2017) demonstrated that a reduction of on-road glance duration impairs hazard detection performance. Glaser et al. (2017) demonstrated that eyes-off-road time negatively impacts driver performance when resuming manual control in critical scenarios. Additionally, gaze can be indicative of cognitive load or distraction (Wang et al., 2014), fatigue and intoxication (Victor et al., 2005).

As drivers' understanding of the automation develops with experience, so will their usage and monitoring behaviour (Sullivan et al., 2016). On road and simulator studies demonstrated substantial differences between drivers with and without driving automation experience. Larsson et al. (2014) compared control transition performance in a simulator between ACC users and drivers without prior driving automation experience and found that while automated driving increased response time, experienced users responded faster than novices in cut-in scenarios. Victor et al. (2018) however demonstrated on a 30 minute test track drive in a Volvo XC90 that expectation mismatch during first-failures can result in a crash even with attentive drivers. Stapel et al. (2019) demonstrated that Tesla owners experienced with automated driving perceive a lower workload during automation use compared to first-time users. However, this perception was contrasted by a slower response time on an auditory detection-response task, indicating an increased objective workload when using automation. Hancock and Matthews (2018) provided reflections on the occurrence of similar dissociations. Large et al. (2019) performed a 5-day longitudinal simulator study on conditional automation (SAE level 3) showing high automation usage, trust and secondary task uptake throughout the drives. Time spent attending the road during automation use reduced from 30% to 20% over the five days accompanied with reduced driving performance after resuming manual control. The latter improved after introducing a routine for regaining situation awareness.

While several studies were conducted in controlled or semi-controlled on-road conditions, only a few have investigated the use of and adaptation to automated driving in a naturalistic setting. Beggiato et al. (2015) performed a longitudinal on-road study where they found that drivers developed their trust and functional understanding of ACC over ten drives while establishing a high acceptance within two drives. Morando et al. (2019) investigated how SAE2 driving automation influences attention during 10 months of naturalistic manual and automated driving by 17 participants. They found longer on-road glances and lower percent eyes on road centre during automated driving (ACC and LK) compared to manual driving. The latter was interpreted as a reduced task demand during automation use. Russel et al. (2018) conducted a naturalistic driving study with 120 participants driving vehicles equipped with adaptive cruise control and automated lane keeping for 4 weeks. They report the effects of traffic stability, road type and weather conditions (no-precipitation vs precipitation) on automation use and found that drivers were performing secondary tasks 60% of the observed time regardless of automation use and found no difference in percentage eyes-off-road time, off-road glance duration or type of secondary task. Reaction times to the 'hold steering wheel'- requests did not change over the four weeks of use, but instances occurred in the first week where such requests were intentionally ignored to investigate the vehicle's response. While these studies provide useful insights, the evolution of behaviours from manual to automated driving has mainly been examined for the first experiences with automation, or lack observations of baseline driving prior to developing experiences with automated driving.

In this study we report preliminary findings on automation use and driver attention from a longitudinal naturalistic driving study conducted in the Netherlands. The study is unique in its inclusion of a one month manual driving baseline followed by a two month experimental phase with the same participants and vehicles where participants were allowed to use the vehicle's automation, enabling a within-subject analysis of behavioural adaptation over the first two months of automation usage.

We addressed the following two research questions:

1. When and where do drivers use ACC and/or lane keeping, as a function of road type and driving speed, time in trip, period of day and automation experience?
2. Is driver attention different during manual driving and supervised automation?

We studied automation use and driver visual attention allocation. In order to perform these analysis, we explored to which extent the visual annotation of automation status and driver attention can be automated. We trained a classifier to identify system icons in the instrument panel using video and to classify driver attention distributions among attentive regions and regions associated with non-driving tasks using head pose estimated from video. Both classifiers were trained and evaluated on manually annotated data from the naturalistic study.

In the following sections, we provide a brief description of the used dataset and methodologies for data preparation, and provide an overview on the data selected after filtering. The results section provides insights in driver's usage of the automation and attention distribution including longitudinal effects.

3.2. METHODS

3.2.1. DATA DESCRIPTION

In a collaborative project conducted by TNO, SWOV and the Dutch ministry of Infrastructure and Watermanagement, the RDW (Dutch Vehicle Authority) and RWS (Dutch Road Authority), recent passenger cars with SAE level 2 automation were equipped with additional instrumentation to observe the driver and the environment. Naturalistic driving data was collected providing these vehicles for daily use to drivers having no prior experience with SAE level 2 automation. The naturalistic dataset is unique in that it includes one month of manual driving (baseline condition) followed by two months of use with automation under naturalistic driving conditions (experimental condition), allowing for a longitudinal within-subjects analysis of how automation use changes over time. The full dataset includes five vehicle types (BMW 540i, Tesla S, Mercedes E, Volkswagen Golf E, Audi A4 Avant) driven by 20 participants. However, the data from only two vehicle types (Tesla and BMW) and 9 participants are currently included in this paper. An overview of the kilometres driven is provided in Table 3.1. Currently 379 trips without automation are compared to 775 trips during which automation was available. For the remaining recordings automation status was either unavailable or inaccessible at the time of this analysis.

Both the BMW and Tesla were equipped with full-range ACC and lane centring. The BMW ACC operated for speeds between 0-180km/h while the Tesla ACC operated between 0-150km/h. In the BMW, lane keeping permits hands off steering wheel for up to 25 seconds. While enabled, the BMW system engages automatically whenever system requirements are met (e.g. clear lane markings) and allows the driver to provide corrective steering without disabling the automation. Tesla lane keeping permits 15 seconds of hands free driving and becomes unavailable for the remainder of a drive when this limit is exceeded 3 times. Tesla's lane keeping has to be engaged by the driver and turns off when the driver provides corrective steering or braking. Lane keeping use with or without ACC enabled. The Tesla only allows lane keeping use while ACC is on.

Table 3.1: Overview of the data collected with one Tesla (three participants) and four BMWs (seven participants in total) equipped with supervised automation functionality.

	Km driven	Days baseline: not using automation	Days experimental: automation available
Tesla1	18 624	47	64
Tesla2	13 713	34	61
Tesla3	5 417	28	31
Total	37 754	109	156
BMW1	14 746	62	59
BMW2	7 887	37	68
BMW4	8 993	40	61
BMW5	7 843	31	61
BMW6	14 638	49	91
BMW7	20 566	42	94
BMW8	11 988	32	75
Total	94 676	321	552

3

PARTICIPANTS

For two participants (1 BMW, 1 Tesla), the demographic data was not available. The remaining 7 participants were all male, mean age 49 years (σ 5.2 years), had their licence for 29.1 years (σ 6.2 years) and had driven 30,000km to 40,000km in the 12 months prior to the experiment. All participants indicated they felt "very interested" and "averagely" to "well" informed about the latest technological developments in the vehicle sector. Prior to the experiment, all but one participant normally used a vehicle equipped with cruise control, zero with adaptive cruise control or lane keeping assistant and three with lane departure warning. One participant (Tesla group) indicated to frequently use a lane keeping assistant.

INSTRUMENTATION

Each vehicle was retrofitted with eight cameras observing the driver, instrument cluster, exterior in forward, left, right and rear directions, pedal bay and a top-down view towards the driver seat. The drivers were observed with a 325x288 resolution at 10Hz. The Tesla instrument panel was observed with a 720x576 resolution at 25Hz. Figure 3.1 provides an overview of the available video feeds.

Figure 3.1: Overview of the eight camera perspectives recorded by the TNO instrumentation in the time-synchronized visualisation by SWOV for each vehicle. In reading order: right mirror view, forward view, left mirror view, driver face, instrument panel, rear view, driver seat, pedal bay. Driver is not occluded in actual footage.

CAN-bus data was collected, from which we derived:

- velocity
- steering wheel angle
- brake and accelerator pedal
- turn indicator
- lights (front and back)
- wind screen wipers
- steering torque
- accelerations (3-axis)
- ACC reference distance and speed*
- ACC and LK status*
- LDW and collision warning*

* retrieval was successful for BMW only.

A smart camera system (MobilEye) was specifically installed for the study, recorded lane position and surrounding road users. For map-matching, GPS and IMU data were collected at 1Hz and 10Hz respectively. All signals except video were time-stamped. Video recordings were not synchronised but were watermarked with a human-readable timestamp before being stored with lossy compression.

RECORDING QUALITY

A number of challenges were unveiled after data collection. Some videos were corrupted and had to be omitted from the analysis. Reverse engineering of CAN bus data identifying automation status or automation interactions (button presses) was successful for the BMW but not for other vehicle types. For the BMW, automation status was available for only 12% of the trips at the time of analysis, limiting the analysis of the BMW to 22 trips from one participant during baseline and 344 trips from 3 participants during the experimental phase. The missing BMW data is being retrieved by TNO and will be analysed in the future. For the Tesla automation status was inferred from camera images of the instrument panel as described below.

3.2.2. DATA PREPARATION

TIME SYNCHRONISATION

While numeric data is time-stamped, videos only tracked the number of frames since the start of a recording. Since the data logger did not start all recordings simultaneously, a delay between first log time from the data logger and first frame of the video occurred, which could be as large as 5 seconds. Dropped frames introduced additional offsets. Inspection of several datasets indicated that the delay due to dropped frames was typically limited to a few seconds per 30 minutes. To correct for the delay between first frame and first log time, timestamp watermarks present in the video were interpreted using template matching, providing the first frame where the watermark increments by a second. While the same method could be used to identify and correct for frame drop, this has not yet been performed at the time of writing. The results reported here have thus been corrected for initial offsets, but assume a constant frame rate without frame drop. To account for the possibility that video derived data (automation status for the Tesla and head pose estimates) may be somewhat out of sync, the analysis avoids timing-critical analysis like reaction time.

AUTOMATION STATUS FOR THE TESLA VEHICLE

For the Tesla, automation system status was retrieved from the instrument cluster videos through icon template matching and a simple neural net classifier. This approach was not needed for the BMW as automation status was retrieved from CAN data. This approach was deemed infeasible for the other vehicles due to poor icon visibility in the recordings (which were challenging to discern even for manual annotation). In the Tesla four icons could be displayed to indicate system status: ACC-on, ACC-available, LK-on and LK-available. Only one icon for ACC and one for LK could be available at any time. System unavailability is communicated by an absence of icons. For each icon, three template images were selected to represent different light conditions and camera perspectives. Maximum confidence values were obtained for each template once every 12 frames (2.1 Hz) through OpenCV template matching performed on a 150 by 245 px subspace to account for camera movement. The required subspace was estimated by overlaying one frame from the middle of each video and visually inspecting where icons occurred. Since logistic regression and manual threshold tuning did not yield accuracies above 70% , these confidence values were presented to a simple neural network consisting of two hidden layers and leaky ReLu activation functions with a 0.1 negative slope. The full network and template icons (relative size differed as depicted) are illustrated in Figure 3.2.

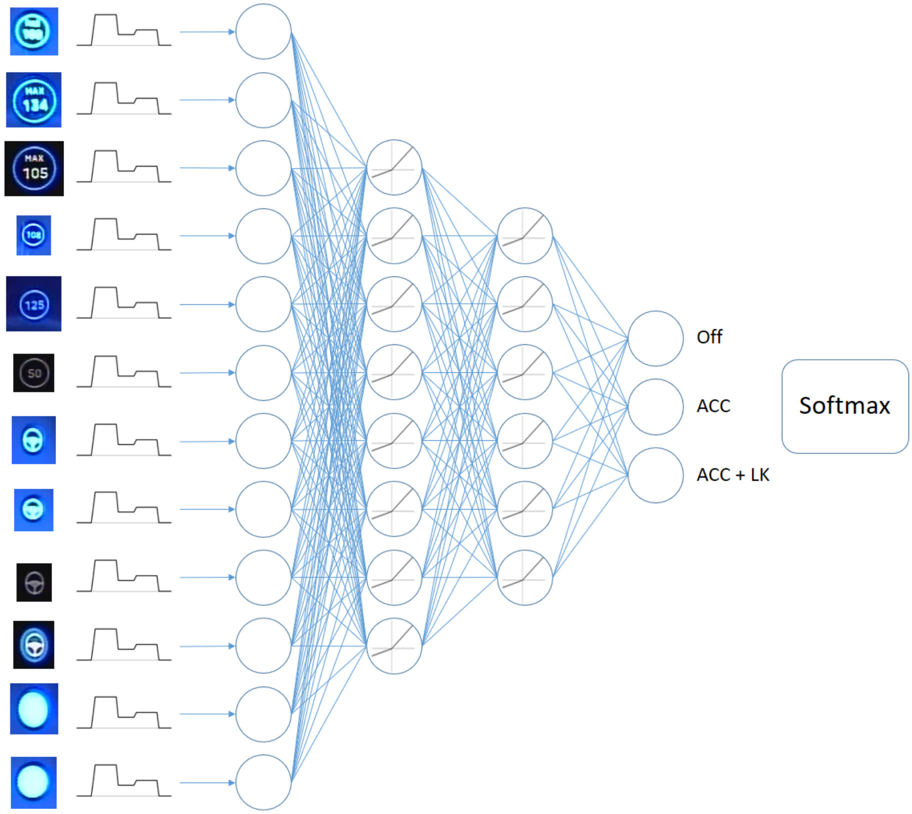


Figure 3.2: Illustration of the classifier setup. Template matching was performed with 12 icon samples. The maximum normalized correlations were used as the input features of a neural network with two hidden layers (leaky ReLu activation functions with a negative slope of 0.1)

To train the classifier, 1628 status transitions were manually annotated among 121 randomly sampled recordings. This resulted in 206,653 frames for training. A test set with 445 transitions was annotated on an additional 27 videos, resulting in 108,396 frames for testing. Annotations were limited to the (disjoint) classes manual, ACC and ACC+LK. System availability was not classified since it would disproportionally increase annotation efforts due to its high transition rate. Classification performance on the test set is shown in Table 3.2 and resulted in an overall accuracy of 99.33% , which was considered sufficient for the current analysis.

Since performance on the test set was used as a stopping criterion for classifier design efforts, performance of the final classifier was verified on another set of 69 randomly sampled videos through visual inspection. Among these, 1342 minutes of manual, 57 minutes of ACC and 257 minutes of ACC+LK use were observed. Approximately 3 minutes (0.2%) worth of frames were misclassified among these videos. Misclassification occurred when the icons were particularly challenging to detect from the video. Common artefacts include rolling image, occluding specular reflection and intensity overflow, as illustrated in Figure 3.3. Specular occlusion typically resulted in momentary misclassification of a single frame. Pixel overflow could last

Table 3.2: Confusion matrix comparing automation status between manual labelling and NN classifier on the test set. Accuracy is 99.33%.

		Predicted		
		Manual	ACC	ACC+LK
Annotated	Manual	54 347	55	100
	ACC	195	10 592	101
	ACC+LK	55	218	42 733

for several seconds but was found to have negligible impact on classification performance. When rolling images occurred, they affected an entire recording. Out of 100 randomly sampled videos, 28 had pixel overflow, and 6 had rolling images.

Figure 3.3: Examples of glitches in video data on which icon recognition may fail. Left: rolling image of the instrument panel as result of lost vertical sync. Middle: specular reflection of a hand occludes the ACC icon. Right: intensity overflow turns over exposed white areas black in the in the LK icon.

Recordings with rolling images resulted in a high status change frequency whenever ACC or ACC+LK was available. Based on spectral power diagrams, the 100 most suspect videos were manually checked for tearing and 45 videos with rolling images were removed from further analysis based on this check.

HEAD MOTION TRACKING & REGIONS OF INTEREST CLASSIFICATION

Dedicated head or eye tracking technology was not considered during instrumentation design. Hence, head motion was tracked using the low-end driver facing camera footage using OpenFace 2.0, an open-source facial behaviour analysis toolkit. It maintains a mean absolute error of 3^{o} under various light conditions and facial expressions (Baltrusaitis et al., 2018). While OpenFace can also estimate eye gaze direction, its intended application is gaze tracking under relatively small angles (laptop screens) and was found to perform poorly on our database. Possible causes are the videos resolution, compression and the angle between the camera and the driver's forward facing direction ($\sim 30^{o}$). Gaze estimation was therefore not extracted. This also reduced computation time to one month for extracting head position and orientation on the current dataset.

Several studies have suggested that head pose can be an acceptable gaze substitute when classifying attention into relevant regions of interest. Lee et al. (2018) have demonstrated that attention classification from head pose is feasible for on-road driving and obtained classification accuracies in the order of 83% and higher. Similarly, Braunagel (2017) used head pose as a fall back for eyes-on-road classification when gaze data was unavailable. Henni et al. (2018) showed that eye based features and head based features can achieve a similar classification performance for on-road drowsiness detection.

We attempted to classify attention allocation in the Tesla to the regions of interest (ROI) defined in Figure 3.4. Regions were selected for their functional purpose during driving; Left, right, windshield and instruments represent regions that are relevant for the driving task while distracted and centre console are not relevant to the driving or monitoring task.

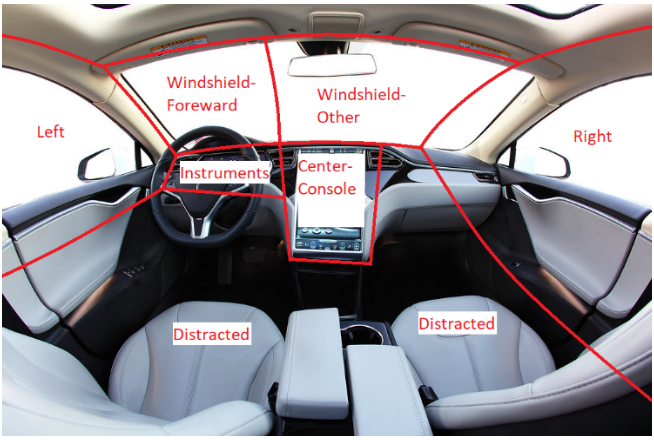


Figure 3.4: Illustration of the approximate head pose regions of interest (ROI).

Since OpenFace estimates head pose relative to the driver-facing camera, association of these head poses to regions of interest requires calibration. Unfortunately, the camera tends to move considerably among recordings, making static ROI definitions impractical. Instead we used typical driver behaviour to correct for camera displacement. A common approach is to create a histogram of all head poses and assume that the distribution modes (the most frequent direction) corresponds to facing the road centre on a straight road. One challenge with this approach is that it does not account for momentary postural changes, which may alter the relation between head pose and gaze direction. To account for this, (Ahlstrom et al., 2012) identified multiple peaks as road-centre facing poses, amongst other refinements. In this paper, we adopted a geometric solution in which the head heading and pitch are compensated for movement of the head's location. While the ideal solution would be to calculate where the facing direction intersects the vehicle's internal geometry, this becomes impossible when the interior's location relative to the camera is not known or changing across drives. We therefore determine the facing direction's intersection with a sphere with a 2 m radius centred at the 50th percentile head origin. This origin is determined for each trip and uses periods of highway driving if available, or all data otherwise. We then express this intersection in polar coordinates to retrieve a heading and pitch compensated for head location. These angles can then be centred to the forward facing reference angle, for instance using the 50th percentile head pose.

This last step has however not yet been performed, which means that the classification analysis has been performed with correction for changes in camera orientation. This correction was applied to the head deviation analysis described in the results section.

To create a ground-truth classification of head poses, we manually labelled 10,552 images from the driver-facing camera into six attentive and distractive regions following the scheme in Annex A. To balance the distribution of samples across regions, frames were sampled to obtain a uniform distribution of head poses. For the second half of the annotations, we filtered to only annotate stationary head poses since we found transitions between regions were often hard to classify. Since only very few poses were labelled as attending the instruments, this class was merged with windshield-forward, with which the samples overlapped best. Figure 3.5 shows a scatter of all annotated head poses for sphere-projected heading and pitch angles.

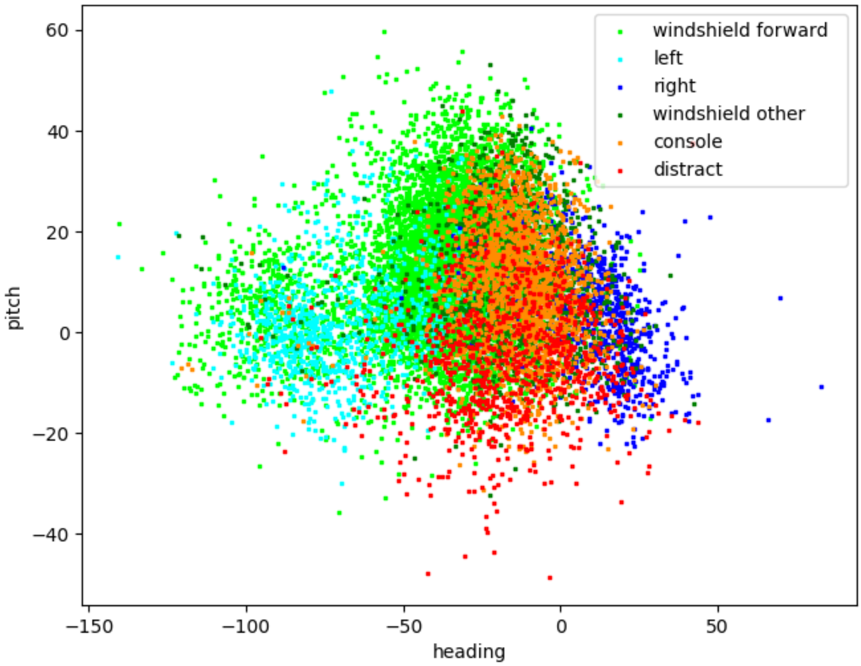


Figure 3.5: Scatter of all annotated head poses. Values are compensated for head location through sphere projection, but orientations have not been corrected for camera placement. Hence zero heading is directed towards the camera and a heading around –35° is forward. Positive pitch represents facing upward.

PERFORMANCE OF ROI CLASSIFICATION

Head pose was classified using a radial basis function support vector machine. 60% of the annotations were used for training and 40% for testing. Table 3.3 shows the confusion matrix of the classifier on the test set. While accuracies between windshield-forward and other regions (Left: 92.7% , console: 86.4% and 95.7%) are similar to those reported by Lee et al. (2018) under similar conditions and methods, our overall accuracy is only 69% . Despite balancing head poses through uniform sampling, windshield forward received 6 times more annotations during manual labelling compared to the other categories on average. As a consequence, the classifier is biased towards this category and inflates accuracy for paired comparisons with that category. The intersection over union rates indicates performance without rewarding true negatives and thus provide a better indication of classification performance per category. Even a binary classification between driving related and unrelated attention does not perform well. Grouping the interior-facing categories (distracted and centre console) results in an intersection over union of 41.3% which is insufficient for reliable distraction identification. The main source of confusion is the ambiguity between head facing direction and direction of gaze, which is especially large in pitch but also in heading for angles further away from road centre.

Due to the disappointing ROI classification performance we will not analyse attention distribution or transitions over these regions as function of automation use, driving environment, and experience in using automation. Instead we will use head pose deviation as an indicator of attention.

Table 3.3: Confusion matrix and intersection over union (IOU) for head pose classification, comparing human annotations to RBF SVM classifier over 4214 test images in the Tesla. Each cell indicates the number of images (top), percentage of ground truth annotated class (bottom left) and percentage of predicted class (bottom right)

		Predicted						
		Windshield forward	Left	Right	Windshield other	Centre console	Distracted other	IOU
Annotated	Windshield forward	2200	30	12	3	74	11	69.8%
	Left	152	124	0	0	1	2	39.9%
	Right	29	0	199	7	68	4	49.5%
	Windshield other	234	1	44	6	87	4	1.5%
	Centre console	318	1	23	11	285	29	29.8%
	Distracted	90	0	16	0	60	89	29.2%

3.2.3. DATA AVAILABILITY

In order to perform our analysis, the various information sources had to be filtered, synchronised and re-sampled. This process is detailed in Annex B and the resulting data availability is described in Table 3.4.

3.3. RESULTS

3.3.1. AUTOMATION USAGE

For automation use during the experimental condition, we first describe the distributions for both the Tesla and BMW and then provide a statistical analysis for the Tesla. Statistics for the BMW will be added in future versions of this manuscript adding data of more participants. Automation status is observed with respect to road type, road speed limit, driving speed, time since the start of a trip and time of day. When a data point misses required information to contribute to a particular image, this data is omitted only for that particular visualisation or analysis.

Table 3.4: Data availability among the trips used for analysis after resampling and filtering.

	Tesla[1]		BMW[2]	
	Data successfully matched with system status (if any)	Proportion of trips with data available	Average data used per trip (if any)	Proportion of trips with data available
Automation status		100%		12.0%[3]
Speed km/h	80.3%	76.0%	99.9%	38.3%
Allowed speed km/h	65.9%	76.0%	80.7%	38.3%
Road type	63.4%	76.0%	75.1%	38.3%
Gearstick	83.1%	76.0%	-	-
Head pose	72.4%	75.5%	95.6%	37.5%

Note1: considers all trips for which automation status has been extracted
Note2: this includes all available trips
Note3: more BMW Automation status data will be retrieved for future versions of this manuscript

During the experimental condition, Tesla users drove manually 55.8% of the time, 8.9% with ACC and 35.3% with ACC+LK. BMW users drove 58.4% manually, 1.2% ACC, 33.7% ACC+LK and 6.7% with LK. Speed limiting was not used by the participants examined so far. Figure 3.6 illustrates automation use by speed limit and road type. For both vehicles, most driving time is spent on the highway, and full automation (ACC+LK) is used most here (Tesla: 63.0% , BMW: 69.9%). Manual driving is however preferred when negotiating highway links. Automation is used very little on roads with speed limits below 70 km/h. In both vehicles, whenever automation is used, preference seems to be towards longitudinal and lateral automation (ACC+LK) over longitudinal only (ACC) or lateral only (manual + LK).

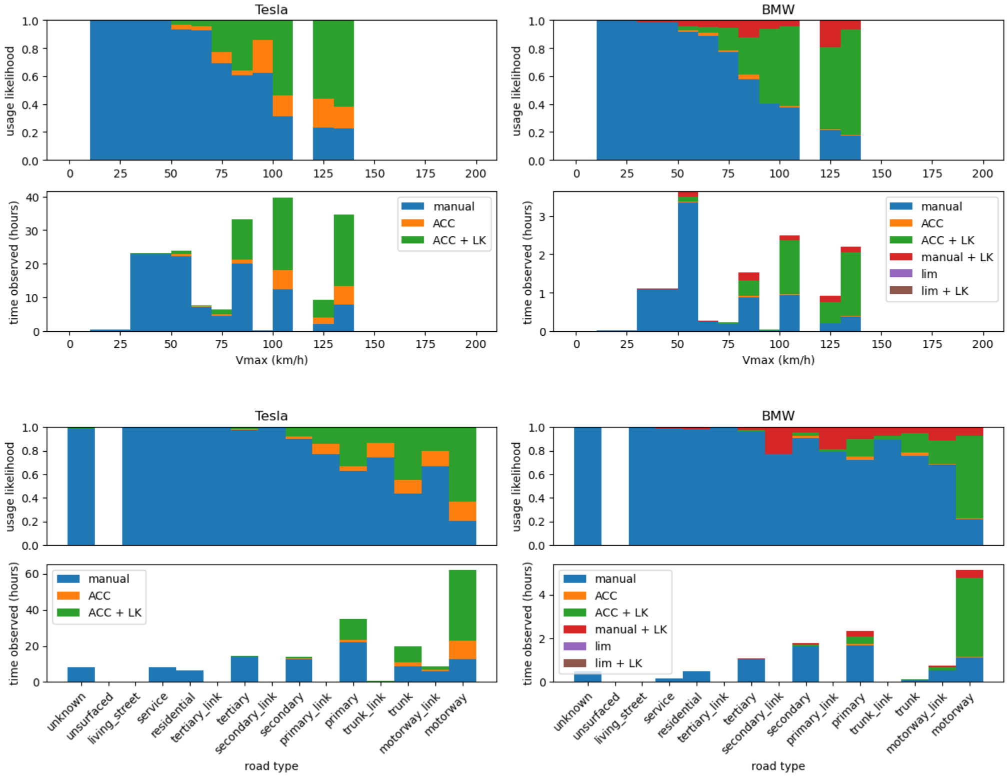


Figure 3.6: Automation use for road speed limit (top) and road type (bottom). Road type was obtained by map-matching using OpenStreetMap. Type descriptions can be found on https://wiki.openstreetmap.org/wiki/Key:highway.

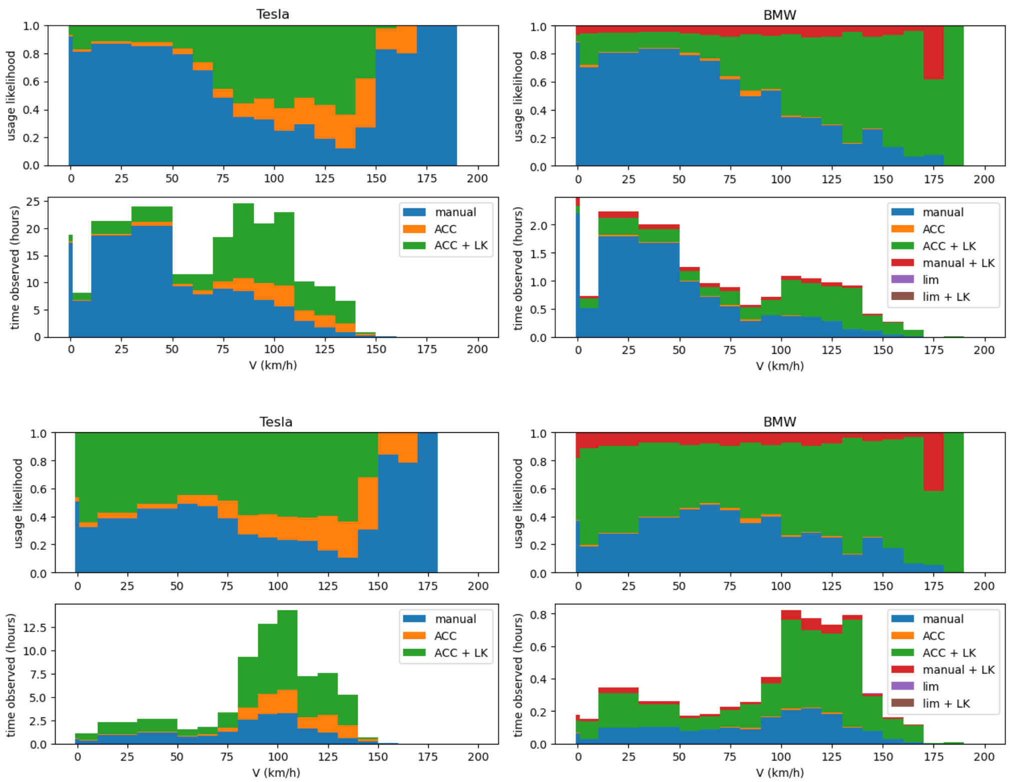


Figure 3.7: Automation use for both vehicle models as a function of vehicle speed for all road types (top) and highways (bottom). Any speed between -1 km/h and 1 km/h is considered stationary. Time where vehicle gear is known to be in “park" is excluded (tesla only).

Figure 3.7 shows how automation use changes with driving speed. Usage is generally low for driving speeds below 70 km/h. However during highway driving, automation use remains high at all speeds, with peak usage during slow stop-and-go traffic (0-30 km/h) and at higher speeds (>80 km/h). At higher speeds a sudden drop in automation use can be observed for the Tesla. This drop corresponds with the upper limit at which the vehicle makes automation available. According to the manuals, the Tesla allows ACC + LK use for speeds up to 150 km/h while the BMW allows ACC use at speeds up to 180 km/h and LK with speeds up to 210 km/h.

Figure 3.8 shows how automation use changes over the duration of a drive. After the first 10-20 minutes, Automation use appears relatively steady. The scatter at later times is an artefact resulting from the low number of trips with long durations. Note that the BMW in particular represents a relatively small number of long trips. Additionally, there is a sharp drop in BMW observations after exactly 30 minutes, suggesting a data availability problem related to the separation of trips into 30 minute recordings. Figure 3.9 shows that automation use also appears to be uniform across the day for both vehicle types.

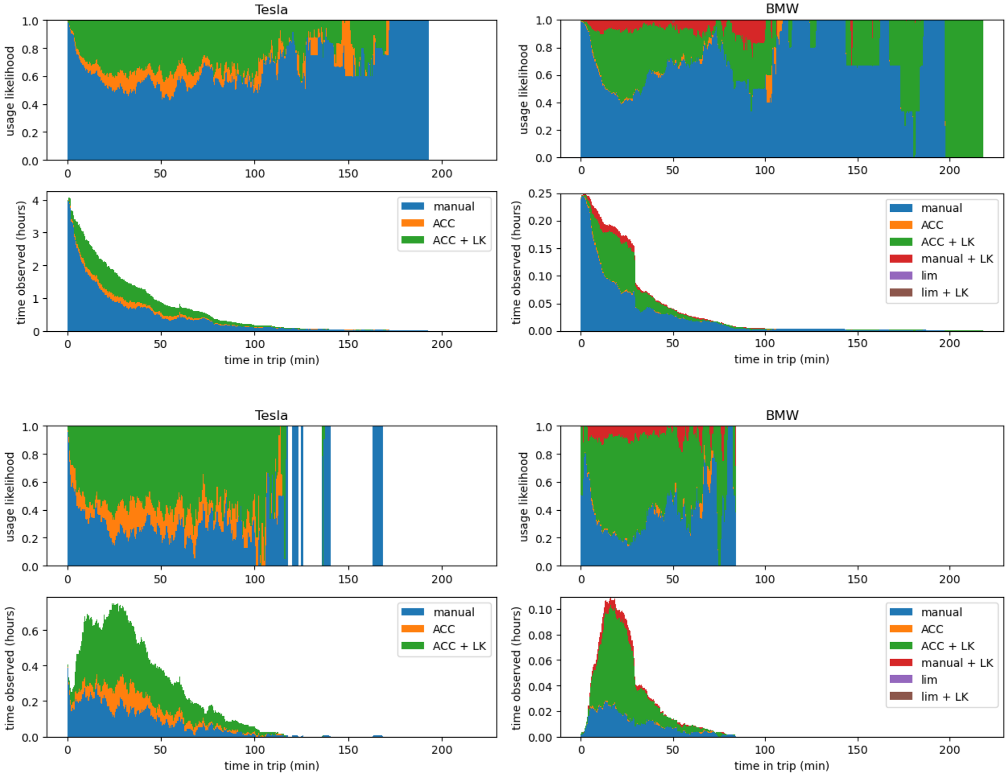


Figure 3.8: Automation usage over time since the start of a trip for all road types (top) and highways (bottom).

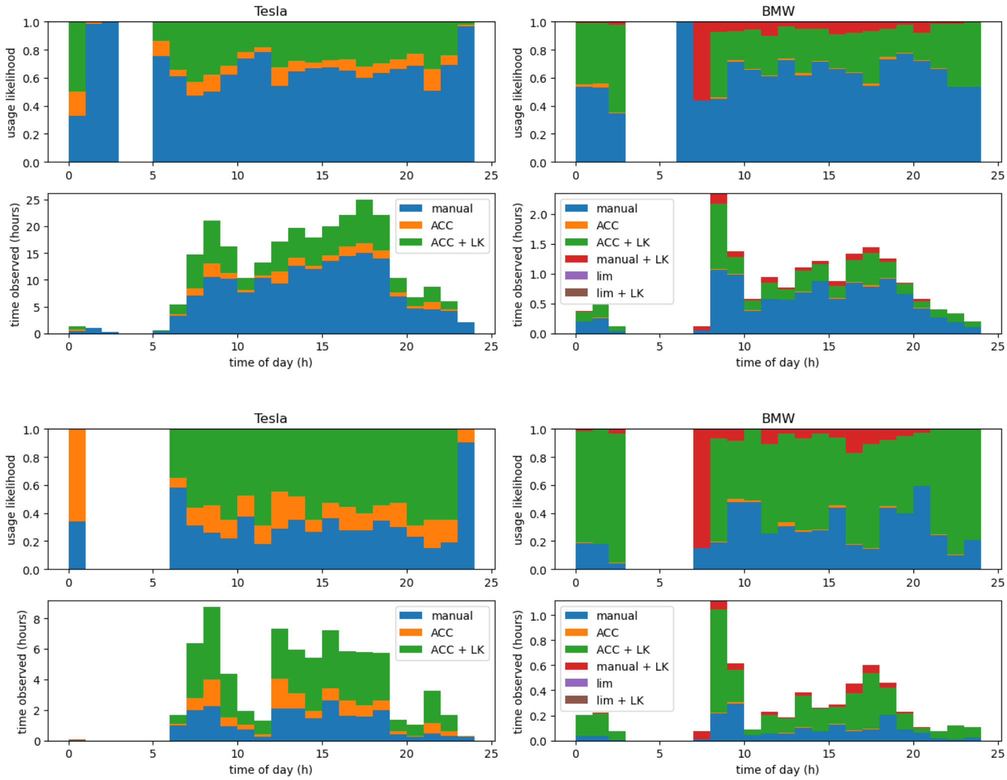


Figure 3.9: Automation use over time of day (Amsterdam DST) for all road types (top) and on highway (bottom).

STATISTICS OF AUTOMATION USE

To evaluate if automation use is influenced by 'time in trip', 'time of day' and 'driving speed', we performed between-trips multilevel ANOVAs with participant as random factor. Only highway driving is considered for this analysis. Since BMW data is incomplete at this time, statistical analyses are only presented for the Tesla. While this means that the analysis represent a small number of participants, it may support the identification of large and consistent effects which should be considered for further analysis. Table 3.5 provides the means and standard deviations for each category and variable. It should be noted that table 3.5 and the histograms of figures 3.9, 3.8 and in particular 3.7 suggest different distributions. This is caused by the histograms showing total usage whereas table 3.5 uses average usage per trip and thus does not account for trip duration. Therefore, the histograms represent exposure whereas table 3.5 is indicative of how often the decision is made to use automation under a given circumstance.

Time of day was split into five categories: night (23:00-4:59 hrs), morning (5:00-9:59 hrs), day (10:00-15:59 hrs), afternoon (16:00-18:59 hrs) and evening (19:00-22:59 hrs). Night time driving was omitted from the statistical analysis due to low sample size. A between-trips ANOVA for time of day while on highway was not significant for ACC+LK use $F(3, 231.3)=0.639$, $p=.591$ or

Table 3.5: Estimated percentage of time automation use on highway; mean and standard deviation over experimental trips

			Time of day (hour)					Time in trip (min)			Speed (km/h)			
			23-4	5-9	10-15	16-18	19-22	0-30	30-60	60-90	0-10	10-60	60-100	>100
Tesla	nr. trips		3	45	97	60	34	212	103	43	95	155	160	155
	Manual	μ	72.9%	47.3%	57.1%	55.3%	47.4%	52.2%	36.5%	38.4%	74.1%	79.0%	42.1%	29.2%
		σ	34.7%	37.8%	37.6%	40.8%	39.7%	38.9%	35.2%	35.0%	40.1%	32.4%	34.1%	33.4%
	ACC	μ	27.1%	10.8%	9.9%	8.2%	10.5%	9.6%	17.7%	13.5%	3.5%	3.9%	13.9%	16.3%
		σ	34.7%	13.7%	14.7%	14.6%	17.5%	15.4%	26.4%	22.5%	16.8%	12.9%	19.9%	20.6%
	ACC+LK	μ	0.0%	41.9%	33.0%	36.5%	42.1%	38.2%	45.8%	48.2%	22.3%	17.1%	44.0%	54.5%
		σ	0.0%	32.9%	31.8%	35.6%	36.0%	34.4%	34.3%	35.4%	37.6%	29.6%	30.9%	33.6%
BMW	nr. trips		19	53	55	50	27	194	60	12	121	191	196	193
	Manual	μ	29.6%	28.5%	39.8%	28.2%	52.2%	33.4%	39.9%	52.8%	56.6%	64.5%	50.5%	27.2%
		σ	29.8%	31.6%	31.4%	35.0%	35.9%	34.2%	34.5%	32.6%	46.7%	42.5%	38.4%	33.1%
	ACC	μ	0.4%	1.2%	1.3%	1.2%	0.2%	0.9%	0.8%	3.2%	0.5%	0.7%	1.4%	0.9%
		σ	0.6%	2.9%	2.9%	2.3%	0.3%	2.3%	2.7%	7.7%	2.8%	4.3%	5.6%	3.0%
	ACC+LK	μ	69.0%	59.9%	53.6%	57.2%	44.8%	57.8%	54.1%	41.2%	31.2%	24.3%	39.5%	64.0%
		σ	29.8%	32.4%	30.0%	36.7%	37.1%	34.6%	34.2%	31.7%	41.8%	36.9%	34.9%	35.4%
	LK	μ	0.8%	10.3%	5.3%	13.5%	2.7%	7.9%	5.2%	2.8%	11.7%	10.5%	8.6%	7.8%
		σ	1.3%	18.7%	10.9%	25.6%	8.2%	18.2%	13.1%	3.4%	27.0%	24.2%	18.2%	20.0%
	lim	μ	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
		σ	0.6%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%
	lim+LK	μ	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
		σ	0.0%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%

ACC use F(3,231.4)=0.312, *p*=.817.

Time in trip was split into three categories of 30 minutes each. A between-trips ANOVA was not significant for ACC+LK use F(2, 354.0)=1.353, *p*=.260. but was significant for ACC use F(2, 355)=5.689 , p=.004. Pairwise comparisons (Table 3.6) show that ACC use was 8% higher (increased from 9.6 to 17.7%) during the second 30 minutes of driving compared to the first 30 minutes, while on the highway.

Table 3.6: Paired time-in-trip comparisons for Tesla ACC usage. Time indicates duration into trip while usage only includes moments of highway driving. See Table 3.5 for descriptives

a - b				Confidence Interval	
ACC	$\Delta\mu$ (a-b)	SE	*p*	5%	95%
0-30 - 30-60 min	-8.1%	2.4%	.001	-12.8%	-3.3%
0-30 min - 60-90 min	-3.8%	3.3%	.253	-10.4%	2.7%
30-60 min - 60-90 min	4.2%	3.6%	.246	-2.9%	11.4%

We divided highway driving speed into the same categories adopted by (Naujoks et al., 2016). A between-trips ANOVA was significant for ACC+LK use F(3, 559.2)=44.587, *p*<.001 as well as on ACC use F(3,561)=18.937, *p*<.001. Paired comparisons revealed that for ACC+LK, all pairs were significant except between 0-10km/h and 10-60km. For ACC use, all pairs were significant except between 0-10km/h and 10-60km/h and between 60-100km/h and >100km/h (Table 3.7). This suggests that both ACC and ACC+LK are used significantly more often with higher speeds while on highways. However, from a duration perspective, the overall ACC+LK usage in figure 3.7 suggests that ACC+LK is used at lower speeds as much as on higher speeds. This suggests that ACC+LK is especially used during longer periods of slow highway driving, and less when such speeds are only reached momentarily.

Table 3.7: Paired highway driving speed comparisons for Tesla automation usage. See Table 3.5 for descriptives.

a - b				Confidence Interval	
ACC+LK	$\Delta\mu$ (a-b)	SE	p	5%	95%
0-10 km/h - 10-60km/h	4.7%	4.2%	.264	-3.5	12.9
0-10 km/h - 60-100 km/h	-22.1%	4.1%	<.001	-30.2%	-12.9%
0-10 km/h - 100 km/h	-32.7%	4.2%	<.001	40.9%	-24.5%
10-60 km/h - 60-100 km/h	-26.8%	3.6%	<.001	-33.8%	-19.7%
10-60 km/h - 100 km/h	-37.3%	3.6%	<.001	-44.5%	-30.2%
60-100 km/h - 100 km/h	-10.6%	3.6%	.004	-3.5%	-17.7%
ACC					
0-10 km/h - 10-60 km/h	-0.3%	2.3%	.886	-4.9%	4.3%
0-10 km/h - 60-100 km/h	-10.3%	2.3%	<.001	-14.9%	-5.8%
0-10k km/h - 100 km/h	-12.7%	2.3%	<.001	-17.3%	-8.1%
10-60 km/h - 60-100 km/h	-10.0%	2.0%	<.001	14.0%	6.0%
10-60 km/h - 100 km/h	-12.4%	2.0%	<.001	-16.4%	-8.4%
60-100 km/h - 100 km/h	-2.4%	2.0%	.236	-6.4%	1.6%

3.3.2. ATTENTION DISTRIBUTION

Since region of interest (ROI) classification was unsuccessful, we evaluated if automation use changed the head pose distribution. While less informative on the quality of monitoring activity, it can indicate when and to which extent automation use results in substantial changes in monitoring behaviour. Head heading and pitch distributions were centred to the 50th percentile of each trip, and the standard deviation was compared across conditions. For the Tesla, statistical differences were explored by comparing head pose deviation during highway driving with a multilevel repeated measures ANOVA using participant as a random factor.

Figure 3.10 shows the head heading distribution. While large heading angles generally occurred less during automated compared to manual driving, this is mainly attributed to road type since heading distributions are more uniform while on the highway. For the Tesla, highway head heading deviation differs significantly between automation use and baseline driving after correcting for individual differences by using participant as a random intercept $F(3,583)=12.243$, $p<.001$. Pairwise comparisons in Table 3.8 reveal that on the highway, head heading deviation during ACC use was significantly smaller than in all other conditions. Heading deviation during ACC+LK did not differ from manual baseline driving, but heading deviation was larger when the driver decided to drive manually during the experimental phase. Head heading deviation during manual driving was larger during the experimental phase compared to baseline.

Figure 3.11 indicates Tesla users tend to face up more and face down less while using automation, whereas BMW users tend to have a wider distribution of pitch angles while using automation compared to manual driving. Highway head pitch deviation differed significantly among automation usage for the Tesla F(3, 581.2)=8.412, p<.001. Pairwise comparisons in Table 3.8 indicate that head pitch deviation during ACC+LK was indistinguishable from manual driving in both the baseline and experimental phase. Pitch deviation was significantly smaller for ACC compared to the other conditions.

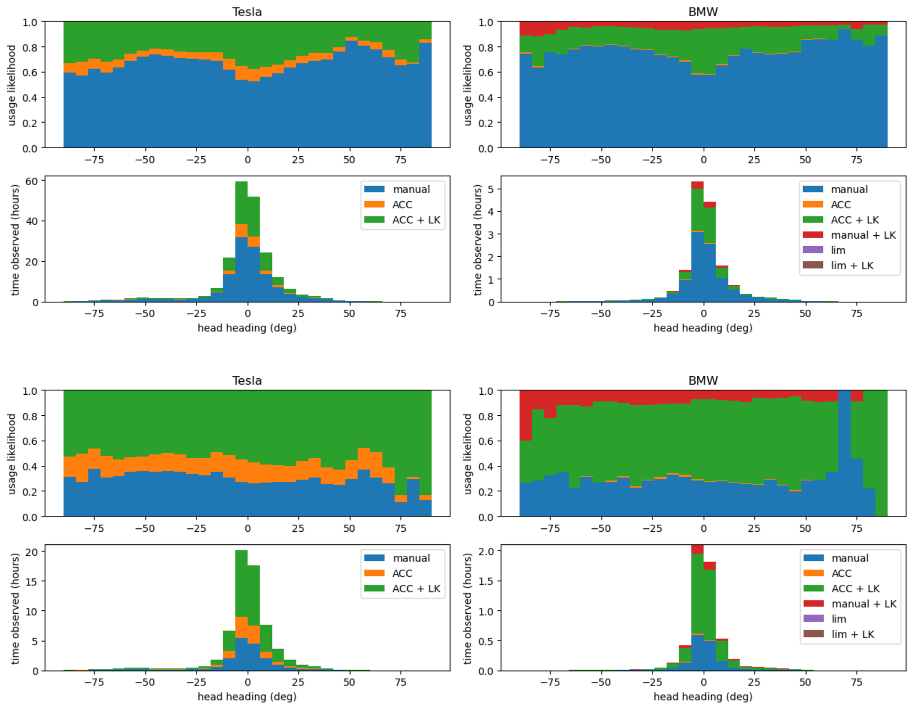


Figure 3.10: Standard deviation of head heading on all road types (top) and on highway (bottom).

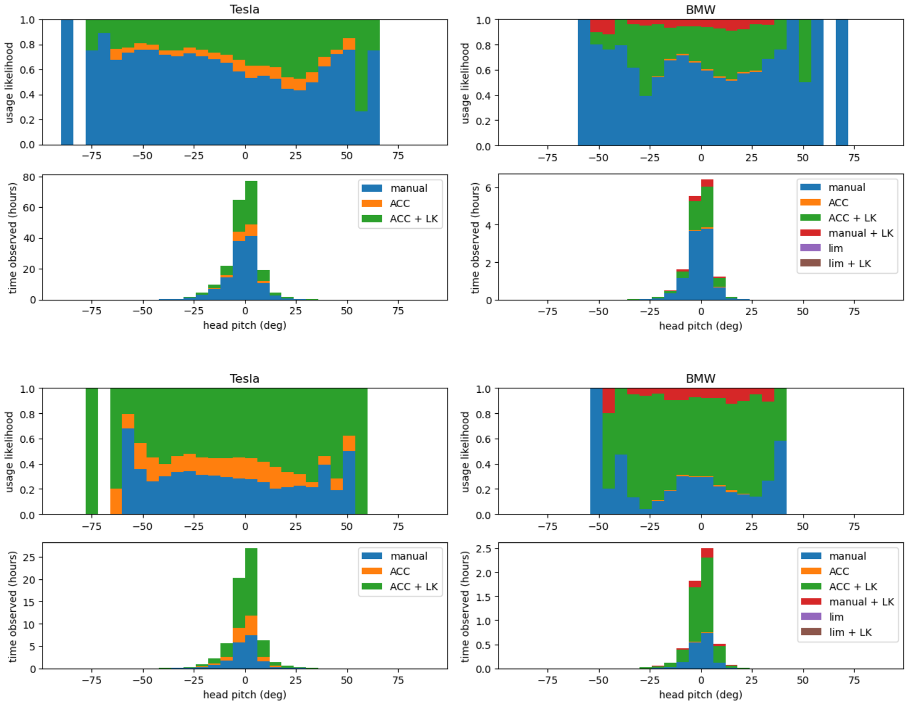


Figure 3.11: Standard deviation of head pitch on all road types (top) and on highway (bottom). Positive pitch is upward.

Table 3.8: Pairwise comparisons of highway head heading and pitch deviation for the Tesla during baseline and experimental conditions (manual, ACC, ACC+LK).

	$\Delta\mu$	SE	p
Heading baseline - Manual	−2.2°	0.6°	.001
Heading baseline - ACC	1.5°	0.7°	.025
Heading baseline - ACC+LK	−0.3°	0.7°	.682
Heading Manual - ACC	3.7°	0.6°	<.001
Heading Manual - ACC+LK	1.9°	0.6°	.003
Heading ACC+LK - ACC	1.8°	0.7°	.008
Pitch baseline - Manual	−0.1°	0.3°	.833
Pitch baseline - ACC	1.1°	0.3°	<.001
Pitch baseline - ACC+LK	0.0°	0.3°	.945
Pitch Manual - ACC	1.1°	0.2°	<.001
Pitch Manual - ACC+LK	0.1°	0.2°	.774
Pitch ACC+LK - ACC	1.1°	0.3°	<.001

3.3.3. EFFECTS OF EXPERIENCE

We evaluate how automation experience changes automation usage, and if experience affects attention as indicated by head pose deviation. The descriptives for automation usage over experience are given in Table 3.9. Overall, automation usage does not suggest any trends over time for the Tesla except for a slight decrease in ACC use, while BMW users seem to use less automation (of any type) over time. Figure 3.12 shows how automation usage varies with time for each road type. It does show a small amount of automation use in residential areas and service roads, though the use is incidental rather than habitual on these road types. In case of the Tesla, ACC+LK also includes use of the auto-park and summon feature. Table 3.10 shows that there are no significant effects of experience on any road type for the Tesla. This suggests that duration wise, automation usage does not change much with experience for the Tesla.

Table 3.9: Descriptives of automation use over experience. Automation use during manual baseline of the Tesla is attributed to misclassification on the video processing.

			Baseline	day 1	Wk1	wk 1-3	wk 4-6	wk 6-9	wk 9-12
Tesla	Nr. trips		401	17	105	256	133	69	30
	manual	μ	99.8%	76.5%	77.4%	79.9%	79.2%	72.8%	83.2%
		σ	1.5%	27.8%	26.5%	24.9%	25.4%	25.8%	26.8%
	ACC	μ	0.1%	7.1%	5.1%	4.3%	4.6%	3.5%	1.6%
		σ	1.0%	10.1%	8.7%	8.9%	8.4%	6.8%	3.9%
	ACC+LK	μ	0.1%	16.4%	17.5%	15.7%	16.2%	23.7%	15.2%
		σ	1.1%	23.2%	22.1%	20.6%	20.2%	22.5%	24.0%
BMW	Nr. trips		22	9	73	150	124	85	4
	manual	μ	100.0%	59.1%	66.9%	67.9%	72.5%	78.6%	84.5%
		σ	0.0%	43.2%	32.6%	31.1%	29.1%	26.5%	23.1%
	ACC	μ	0.0%	0.5%	1.5%	1.3%	0.5%	0.5%	0.5%
		σ	0.0%	1.2%	3.9%	3.7%	1.4%	1.3%	0.9%
	ACC+LK	μ	0.0%	15.5%	22.2%	24.3%	22.7%	17.2%	13.6%
		σ	0.0%	24.3%	26.8%	26.3%	25.2%	22.7%	19.6%
	LK	μ	0.0%	24.8%	9.4%	6.5%	4.3%	3.1%	1.4%
		σ	0.0%	28.5%	18.2%	14.0%	9.1%	7.4%	2.8%
	lim	μ	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%	0.0%
		σ	0.0%	0.0%	0.0%	0.0%	0.1%	6.1%	0.0%
	lim+LK	μ	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%
		σ	0.0%	0.5%	0.2%	0.1%	0.0%	0.0%	0.0%

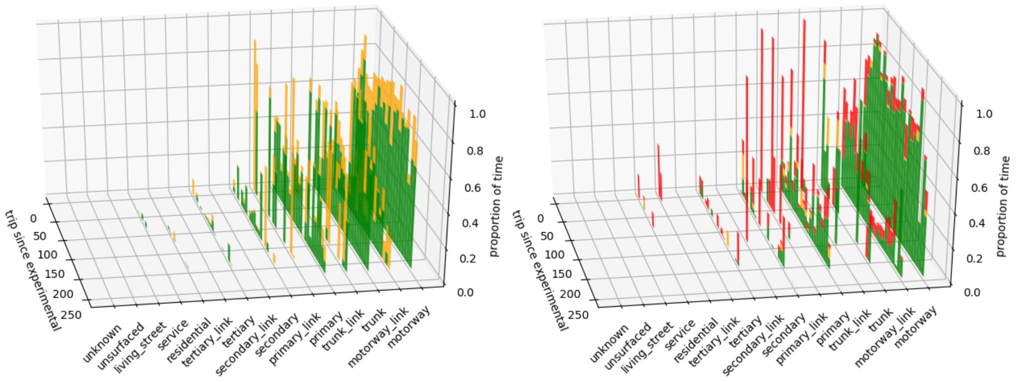


Figure 3.12: Automation usage for Tesla (left) and BMW (right) over experience (trip since start of experimental condition) for various road types. Green=ACC+LK, yellow=ACC, red=LK, (purple=lim and black=lim+LK). Manual driving represents the remainder proportion.

Table 3.10: Between-trips ANOVA with participant as random factor indicating effect of experience on automation usage for the Tesla. Experience is represented by 4 bins: wk 1-3, wk 3-6, wk 6-9, wk 9-12. N=number of trips included in analysis. *p*-values are not corrected for multiple comparisons.

		ACC+LK		ACC	
	N	F	*p*	F	*p*
Motorway	160	F(3,114.3)=1.001	.395	F(3,156)=1.196	.313
Motorway link	158	F(3,154)=0.982	.403	F(3,109.5)=0.324	.808
Trunk	117	F(3,112.2)=0.620	.604	F(3,112.6)=0.286	.836
Trunk link	54	F(3,44.7)=0.119	.948	F(3,46.9)=0.855	.471
Primary	270	F(3,243.8)=0.114	.952	F(3,260.4)=0.400	.753
Secondary	201	F(3,197)=0.840	.473	F(3,197)=0.606	.612

Attention was affected by experience, as inferred from head pose variance. We analyse highway driving for the Tesla only. Effects were corrected for individual differences by using participant as a random intercept. Table 3.11 gives the main effects for head heading and pitch deviation while on the highway and Table 3.12 provides the pairwise comparisons. The 30 trips in weeks 10 to 12 are all made by one participant.

During manual highway driving, head heading deviation was higher throughout the experimental condition compared to baseline, but did not change significantly over time within the experimental condition. During ACC use, head heading was not affected by experience. During ACC+LK use, head heading deviation was not affected significantly by experience at a 5% confidence, though there is a tendency to increase with experience over the first 9 weeks of experience.

During manual driving, deviation in head pitch did not change over time and remained the same as for baseline. Pitch deviation over experience was also not statistically significant during ACC use, though it tended to increase over the first 6 weeks of automation use. A significant trend in the same direction is observed during ACC + LK use.

Table 3.11: Main effects of experience (baseline, wk 1-3, wk 4-6, wk 7-9, wk 10-12) on head heading and pitch deviation for manual, ACC and ACC+LK highway driving with the Tesla. ANOVAs are corrected for individual differences. N represents the number of trips included in the analysis. Baseline condition is only included for manual driving

Head heading	N	F	p
Manual	322	F(4, 309.0) = 3.659	0.007
ACC	133	F(3, 129) = 2.032	0.113
ACC+LK	132	F(3,121.6) = 2.528	0.061
Head pitch			
Manual	322	F(4, 316.1) = 1.329	0.259
ACC	133	F(3, 121) = 2.206	0.091
ACC+LK	132	F(3, 128.0) = 7.611	<.001

3.4. DISCUSSION

In this study, we analysed how SAE2 automation use differs between road types and driving conditions. We investigated how attention, as reflected by head pose deviation, differs between manual and automated driving, and we explored how experience with automation changes automation usage and attention over the first two months of use. We further presented our methodology and performance for two data enrichment methods on the naturalistic dataset. We emphasise that the results are preliminary findings obtained from six participants for the descriptives and only three participants for the statistical analysis. The results should therefore not yet be regarded as the definitive conclusions of this study, but rather as a direction for further analysis of the L2 naturalistic dataset. Such further analyses will be pursued before publication as Journal paper.

3.4.1. AUTOMATION USE

Across all road types, automation usage increases with speed for both vehicle types, with more than 92.5% manual driving at speeds below 70km/h for the Tesla and 88.6% for the BMW. The use of map matching for road type classification allowed us to analyse automation use per road type and discriminate low speed driving in congested highways from low speed driving related to the allowed maximum speed. Automation is used most on highways, where it is used across all speeds including slow highway driving, but less during short periods of slow driving. This suggests that users were generally comfortable using these systems during most highway traffic conditions. Further analysis on automation use during slow highway driving is required to explain this behaviour. In terms of exposure time, highway usage occurred the least near speeds of 50 km/h (Tesla: 44.7% , BMW: 45.2%). ACC+LK appears to be used more than ACC for all road types. For the BMW, ACC was used less compared to the Tesla, while usage of LK without ACC was active 6.7% of the time and did not differ proportionately between road types or driving speeds. These findings suggest automation is mostly used on road types for which the systems are intended, with very little time spent using the automation on roads with a speed limit of 50 km/h or less. Automation use on urban roads was limited and incidental, which suggests that users are aware of the system's general limitations and typically act accordingly. This urban use should be examined further with an analysis of activation

Table 3.12: Pairwise comparisons of experience on head heading and pitch deviation for manual, ACC and ACC+LK highway driving with the Tesla. Baseline only contains manual driving.

	Manual			ACC			ACC+LK		
Head heading	$\Delta\mu$	SE	p	$\Delta\mu$	SE	p	$\Delta\mu$	SE	p
Wk1-3 - baseline	1.9°	0.7°	.006						
Wk4-6 - baseline	3.0°	1.2°	.011						
Wk7-9 - baseline	2.9°	1.1°	.008						
Wk10-12 - baseline	0.8°	2.2°	.716						
Wk4-6 - Wk1-3	1.1°	1.2°	.367	3.9°	1.7°	.022	1.4°	1.3°	.292
Wk7-9 - Wk1-3	1.0°	1.1°	.370	1.6°	1.4°	.255	2.8°	1.1°	.009
Wk10-12 - Wk1-3	−1.1°	2.2°	.605	−1.3°	4.5°	.767	−0.7°	2.6°	.786
WK7-9 - Wk4-6	0.1°	1.4°	.961	−2.3°	2.0°	.253	1.5°	1.5°	.308
Wk10-12 - Wk4-6	−2.2°	2.4°	.353	−5.2°	4.7°	.272	−2.1°	2.8°	.459
Wk10-12 - WK7-9	−2.1°	2.3°	.364	−3.0°	4.6°	.525	−3.5°	2.6°	.183
Head pitch									
Wk1-3 - baseline	−0.2°	0.3°	.571						
Wk4-6 - baseline	0.7°	0.5°	.136						
Wk7-9 - baseline	0.5°	0.4°	.270						
Wk10-12 - baseline	−0.7°	0.9°	.432						
Wk4-6 - Wk1-3	0.9°	0.5°	.073	1.5°	0.6°	.018	1.5°	0.5°	.004
Wk7-9 - Wk1-3	0.7°	0.5°	.153	0.8°	0.5°	.136	1.8°	0.4°	<.001
Wk10-12 - Wk1-3	−0.5°	0.9°	.550	0.2°	1.7°	.915	1.2°	1.0°	.232
WK7-9 - Wk4-6	−0.2°	0.6°	.712	−0.7°	0.7°	.335	0.3°	0.6°	.580
Wk10-12 - Wk4-6	−1.4°	1.0°	.147	−1.3°	1.7°	.459	−0.3°	1.1°	.779
Wk10-12 - WK7-9	−1.2°	1.0°	.215	−0.6°	1.7°	.720	−0.6°	1.0°	.542

attempts, since experimental use in unsuitable conditions may be of short duration even when attempted frequently, and thus contribute little to the time-based usage statistics analysed in this study. In particular, successful and unsuccessful automation engagement attempts could be analysed, though this would require further efforts in CAN bus extraction or video annotation. Furthermore, manual inspection of driver attention during these episodes may indicate if these systems are used responsibly in these situations.

For the Tesla, no significant time in trip, time of day or experience effects were found for automation usage, except for an 8% increase in ACC use during the second 30 minutes of highway driving compared to the first 30 minutes of each drive. However the effect is small compared to its variance and not consistent across all bins. Since the data only represents three participants, the effect may be a consequence of individual differences in ACC usage and typical trip length. Data from more participants is needed to evaluate if these findings can be regarded as generalizable. For this, the BMW data will be analysed once available to study effects of time on task, time of day and experience.

3.4.2. HEAD POSE DEVIATION DISTRIBUTIONS

While we were unable to classify head pose into attentive driving related and distracted driving unrelated areas, the analysis of head pose variance did provide some interesting insights.

On the highway, ACC+LK did not differ from baseline manual driving in terms of head heading or head pitch deviation, but heading deviation was larger during manual driving in the experimental phase. While the increase in heading deviation during manual highway driving may be a carryover effect caused by automation use, such effect has to develop quickly since no longitudinal change was found in manual head heading deviation, and would require an explanation how carryover can result in increased activity during manual driving whereas ACC+LK and ACC use suggest either no change or a decrease in heading deviation. It is more likely the result of strategic automation use: drivers may simply prefer to drive manually in situations which require more head deviation, such as when changing lanes. This hypothesis could be tested through selection of lane changes from the available Mobileye data or selection of highway trips from the experimental condition where automation was not used at any point. Important to note is that the effect of introducing ACC+LK depends on whether it is compared against baseline-manual (no difference in heading deviation) or experimental-manual (ACC+LK reduces heading deviation). This may raise caution for studies which compare attention between manual and automated driving without providing an inexperienced manual baseline.

Comparing to baseline driving, ACC resulted in a reduced head heading and pitch deviation while ACC+LK did not. However, head pitch deviation increased with experience for ACC+LK, and while not statistically significant, the same trend was observed for ACC as well as for heading deviation with effect sizes in the order of 1 or 2 degrees over the first 6 weeks. This suggests that head deviation was initially lower during ACC+LK use compared to baseline, but restored to normal as the participants became more experienced. These trends indicate at behavioural adaptation in driver attention during automation use. This adaptation with experience agrees with Kraft et al. (2018), who found that participants with prior ACC experience spent less time looking away from the road compared to automation novices in a simulator study on visual secondary tasks during SAE2 automation.

Collectively, these findings suggest that while initially altered, the amount of attentional activity in terms of head pose deviation may be similar between ACC+LK use and baseline manual driving. This differs from Morando et al. (2019), who found that the median percent at road centre of glances was 3% smaller during SAE2 compared to manual driving. Possible explanations for this difference include the used metrics (gaze vs. head pose), not controlling for periods of following a lead vehicle (which increased percent road centre by 4% during automation use for Morando et al.), and individual differences from the low number of participants presently analysed.

Whether the lower head pose deviation during ACC and initial ACC+LK should be interpreted as an increase or decrease in monitoring intensity remains to be investigated. If drivers were mostly monitoring attentively during automation, increased attentional demand as found in chapter 2 could be used to interpret lower deviation as an increase in attention to road centre or cognitive narrowing due to an increased mental demand. However, it can also be caused by cognitive load from driving-unrelated thoughts (Victor et al., 2005; Wang et al., 2014), a reduced perceived need for visual scanning, or an increase in mind wandering (He et al., 2011). Even when gaze had been obtained in addition to head pose, identification of the correct cause may be challenging since even for gaze dispersion it is not certain if a wider deviation represents more distraction or a better monitoring strategy (Grüner et al., 2017). Classification of attention to driving related and unrelated areas may provide better insights. While automation of such classification was not successful in the present study, the observation that automation changes behaviour over time provides a motivation for further investigation.

3.4.3. Data enrichment

We end this discussion with a reflection on our methods for data enrichment. We demonstrated that extracting automation status from instrument panel recordings can be an effective solution when this information cannot be extracted from CAN messages, provided that the icons of interest are easily discernible in the recordings. The two layer neural network improved classification accuracy from 70% obtained from logistic regression to 99.33%, while template matching prevented the need for training a complete convolutional neural network which likely would have required a larger training and test set to be annotated manually.

Classifying attended region from head pose proved more challenging. While we obtained similar per-class accuracies as reported by Lee et al. (2018), the overall accuracy of 69% and intersection over union metrics smaller than 50% indicated that head pose is not sufficient for attention classification. While classification performance may improve when preventing camera movement or by including head dynamics in addition to poses, we recommend to include eye tracking to infer attention allocation in future studies. Care should also be taken when using manual labelling of attention allocation, since inter-rater agreement was recently found to be higher than actual classification accuracy in a similar setting (Jansen et al. 2020).

3.5. Conclusions

Automation is mostly used on road types for which the systems are intended. On highways ACC+LK was used 63% of time in the Tesla and 70% of time in the BMW. On roads with speed limits below 70 km/h, all forms of automation combined were used less than 8% of the time, and use on urban roads was incidental rather than habitual, which suggests that users are aware of the system's general limitations and typically act accordingly. Head pose analysis indicated visual scanning to be similar for driving with ACC+LK and baseline manual driving, though effects of experience suggest the amount of scanning during automation use to increase over the first six weeks of use.

Scanning activity was larger during manual driving in the experimental condition compared to manual driving over a one month baseline prior to experiencing automation. This observation may have implications for studies that evaluate attention in automation with automation-experienced users, without automation-inexperienced baseline.

We further demonstrated that feature matching combined with a simple neural network can be effective for extracting automation status from instrument panel recordings. We also demonstrated that head pose without information of gaze direction may be insufficient for region of attention classification.

Based on our preliminary findings, we recommend further analysis on activation attempts and monitoring behaviour during automation use in urban conditions. The identified changes in monitoring activity warrants further investigation on the quality of driver attention and distraction during manual and automated highway driving.

This work will be continued on the remaining participants when this data becomes available. Recommended next steps include:

- Examine engagement attempts and monitoring strategies during urban automation use.
- Examine difference automation use between long and short lasting periods of slow highway driving.
- Interactions of automation use on Circadian and other time effects on driver attention are unlikely to be uncovered from the present dataset, unless automation status can be recovered for the other vehicle types.
- Assess if identified head pose differences during highway automation indicate better or worse attention strategies.
- Test the proposition that attention strategies changed during initial weeks of ACC+LK use but later recovered.
- Test if the increase in head pose during manual driving is a consequence of strategic automation use.

Acknowledgment Funding: This work was partially supported by the NWO-TTW Foundation, the Netherlands, under the project "From Individual Automated Vehicles to Cooperative Traffic Management - Predicting the benefits of automated driving through on-road human behavior assessment and traffic flow models (IAVTRM)" -STW#13712.

3.6. REFERENCES

Ahlstrom, C., Victor, T., Wege, C. and Steinmetz, E. (2012) 'Processing of Eye/Head-Tracking Data in Large-Scale Naturalistic Driving Data Sets', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, No. 2, pp.553–564. https://doi.org/10.1109/TITS.2011.2174786

Baltrusaitis, T., Zadeh, A., Lim, Y.C. and Morency, L.-P. (2018) 'OpenFace 2.0: Facial behaviour Analysis Toolkit' in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, pp.59–66. https://doi.org/10.1109/FG.2018.00019

Banks, V.A., Eriksson, A., O'Donoghue, J. and Stanton, N.A. (2018) 'Is partially automated driving a bad idea? Observations from an on-road study', *Applied ergonomics*, Vol. 68, pp.138–145. https://doi.org/10.1016/j.apergo.2017.11.010

Beggiato, M., Pereira, M., Petzoldt, T. and Krems, J. (2015) 'Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 35, pp.75–84. https://doi.org/10.1016/j.trf.2015.10.005

Braunagel, C. (2017) *Ensuring the Take-Over Readiness of the Driver Based on the Gaze Behavior in Conditionally Automated Driving Scenarios.* Dissertation, Tübingen. http://dx.doi.org/10.15496/publikation-23076

de Winter, J.C.F., Happee, R., Martens, M.H. and Stanton, N.A. (2014) 'Effects of adaptive cruise control and highly automated driving on workload and situation awareness, a review of the empirical evidence', *Transportation Research Part F: Traffic Psychology and Behaviour*,

3

Vol. 27, B, pp.196–217. https://doi.org/10.1016/j.trf.2014.06.016

Dijsselbloem, J.R.V.A., van Asselt, M.B.A., Zouridis, S. and Verheij, C.A.J.F. (2019) *Wie stuurt? Verkeersveiligheid en automatisering in het wegverkeer.* https://www.onderzoeksraad.nl/nl/media/attachment/2019/11/28/wie_stuurt_verkeersveiligheid_en_automatisering_in_het_wegverkeer.pdf.

Farah, H., Bhusari, S., van Gent, P., Babu, F.A.M., Morsink, P., Happee, R. and van Arem, B. (2020) 'An Empirical Analysis to Assess the Operational Design Domain of Lane Keeping System Equipped Vehicles Combining Objective and Subjective Risk Measures', *IEEE Transactions on Intelligent Transportation Systems*, pp.1–10. https://doi.org/10.1109/TITS.2020.2969928

Glaser, Y., Glaser, D., Green, C., Llaneras, R.E. and Meyer, J. (2017) 'Driver Hazard Detection and Avoidance Performance as a Function of Eyes-Off-Road Interval Under Partially Automated Driving', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61, No. 1, pp.1914–1918. https://doi.org/10.1177/1541931213601959

Grüner, M. and Ansorge, U. (2017) 'Mobile eye tracking during real-world night driving: A selective review of findings and recommendations for future research', *Journal of Eye Movement Research*, Vol. 10, No. 2, pp.1–18. https://doi.org/10.16910/jemr.10.2.1

Hancock, P.A. and Matthews, G. (2018) 'Workload and Performance: Associations, Insensitivities, and Dissociations', *Human factors*. https://doi.org/10.1177/0018720818809590

Harms, I.M., Bingen, L. and Steffens, J. (2020) 'Addressing the awareness gap: A combined survey and vehicle registration analysis to assess car owners' usage of ADAS in fleets', *Transportation Research Part A: Policy and Practice*, Vol. 134, pp.65–77. https://doi.org/10.1016/j.tra.2020.01.018

He, J., Becic, E., Lee, Y.-C. and McCarley, J.S. (2011) 'Mind wandering behind the wheel: Performance and oculomotor correlates', *Human factors*. Vol. 53, No. 1, pp.13-21. https://doi.org/10.1177/0018720810391530

Henni, K., Mezghani, N., Gouin-Vallerand, C., Ruer, P., Ouakrim, Y. and Vallières, É. (2018) 'Feature selection for driving fatigue characterization and detection using visual- and signal-based sensors', *Applied Informatics*, Vol. 5, No. 1, p.1. https://doi.org/10.1186/s40535-018-0054-9

Jamson, A.H., Merat, N., Carsten, O.M.J. and Lai, F.C.H. (2013) 'Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions', *Transportation Research Part C: Emerging Technologies*, Vol. 30, pp.116–125. https://doi.org/10.1016/j.trc.2013.02.008

Jansen, R.J., van der Kint, S.T. and Hermens, F. (2020) 'Does agreement mean accuracy? Evaluating glance annotation in naturalistic driving data', *Behavior research methods*, https://doi.org/10.3758/s13428-020-01446-9

Kraft, A.-K., Naujoks, F., Wörle, J. and Neukum, A. (2018) 'The impact of an in-vehicle display on glance distribution in partially automated driving in an on-road experiment', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 52, pp.40–50. https://doi.org/10.1016/j.trf.2017.11.012

Large, D.R., Burnett, G., Salanitri, D., Lawson, A. and Box, E. (2019) 'A Longitudinal Simulator Study to Explore Drivers' Behaviour in Level 3 Automated Vehicles' in *the 11th International*

Conference, ACM Press, New York, New York, USA, pp.222–232. https://doi.org/10.1145/3342197.3344519

Larsson, A.F.L., Kircher, K. and Andersson Hultgren, J. (2014) 'Learning from experience: Familiarity with ACC and responding to a cut-in situation in automated driving', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 27. https://doi.org/10.1016/j.trf.2014.05.008

Lee, J., Muñoz, M., Fridman, L., Victor, T., Reimer, B. and Mehler, B. (2018) 'Investigating the correspondence between driver head position and glance location', *PeerJ Computer Science*, Vol. 4, No. 2, e146. https://doi.org/10.7717/peerj-cs.146

Morando, A., Victor, T. and Dozza, M. (2019) 'A Reference Model for Driver Attention in Automation: Glance Behavior Changes During Lateral and Longitudinal Assistance', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 8, pp.2999–3009. https://doi.org/10.1109/TITS.2018.2870909

Naujoks, F., Purucker, C. and Neukum, A. (2016) 'Secondary task engagement and vehicle automation – Comparing the effects of different automation levels in an on-road experiment', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 38, pp.67–82. https://doi.org/10.1016/j.trf.2016.01.011

Park, H., Gao, S. and Samuel, S. (2017) 'Modeling Effects of Forward Glance Durations on Latent Hazard Detection', *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2663, pp.90–98. https://doi.org/10.3141/2663-12

Russel, S.M., Blanco, M., Atwood, J., Schaudt, W.A., Fitchett, V.L. and Tidwell, S. (2018) *Naturalistic Study of Level 2 Driving Automation Functions* (Report No. DOT HS 812 642). Washington, DC: National Highway Traffic Safety Administration.

SAE International (2018) *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, 2nd ed. https://doi.org/10.4271/J3016_201806

Stapel, J., Mullakkal-Babu, F.A. and Happee, R. (2019) 'Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 60, pp.590–605. https://doi.org/10.1016/j.trf.2018.11.006

Strickland, D.L. (2013) *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices*, 78th ed., National Highway Traffic Safety Administration, Department of Transportation (DOT).

Sullivan, J.M., Flannagan, M.J., Pradhan, A.K. and Bao, S. (2016) 'Literature Review of Behavioral Adaptations to Advanced Driver Assistance Systems', *AAA Foundation for Traffic Safety*.

Victor, T.W., Harbluk, J.L. and Engström, J.A. (2005) 'Sensitivity of eye-movement measures to in-vehicle task difficulty', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 8, No. 2, pp.167–190. https://doi.org/10.1016/j.trf.2005.04.014

Victor, T.W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F. and Ljung Aust, M. (2018) 'Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel', *Human factors*, Vol. 60, No. 8, pp.1095–1116. https://doi.org/10.1177/0018720818788164

Vlakveld, W.P. (2019) *Veiligheidseffecten van rijtaakondersteunende systemen. Bijlage bij het convenant van de ADAS Alliantie*, SWOV - institute for road safety research, report R-2019-14.

Wang, Y., Reimer, B., Dobres, J. and Mehler, B. (2014) 'The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 26, pp.227–237. https://doi.org/10.1016/j.trf.2014.08.003

3

ANNEX A: CODE BOOK FOR ANNOTATING HEAD ORIENTATION IN THE TESLA

If frame suitable for annotation:

- Is driver present, and is driver's facing direction clear? No → Space
- Is another person's face clearly visible? Yes → Space
- Is driver's head pose hard to classify or exceptional/unconventional?∗ Yes → Space

If driver faces away:

- Is driver facing or glancing through left window or to left mirror? Yes → Left
- Is driver facing or glancing through right window? Yes → Right
- Is driver clearly looking away from any exterior view, mirror or vehicle display?∗∗ Yes → Distracted

If driver faces forward:

- Does driver glance to instrument panel? Yes → Instruments
- Does driver face/glance well below instrument panel? Yes → Distracted
- Does driver glance towards the rear view mirror? Yes → Windshield other
- Otherwise: Windshield forward

If driver faces towards the camera:

- Does driver glance (just) above the camera's origin? Yes → Windshield other
- Does driver glance at, slightly below or slightly right of camera?∗∗ Yes → Centre console
- Does driver glance slightly left and slightly below camera? Yes → Distracted
- Does driver appear to glance through right half of windshield or right mirror? Yes → windshield other

∗ Examples include sneezing, being mid-motion and severe head tilt
∗∗ When holding nomadic device, consider the direction of attention rather than the activity.

ANNEX B: DATA AVAILABILITY

In order to perform our analysis, the various information sources had to be filtered, synchronised and re-sampled. The information can be classified into three categories, which differ in how time is referenced. Raw numeric data (vehicle speed, BMW automation status) is indexed by *trip*, which can span multiple *drives*, provided they start within 5 minutes of a previous drive. The data of each trip is split into 30 minute *recordings*. Each data point has a time stamp, consisting of a midnight component (UNIX start of day in ms) and a logtime (Unix time since midnight in ms). When time passes into the next day during a trip, logtime is kept monotonically increasing with values beyond 24h. Midnight values are kept constant for the duration of a recording, but will update at the start of the next recording, which means that its value may change within a trip, but not necessarily at midnight. Correct time stamps can thus only be obtained by adding a trip's first midnight (possibly found in a different recording) to a datum's logtime. Video derived data (Tesla automation status and driver head activity) have no time stamps readily available beyond frame numbers, but are approximated by assuming a fixed frame rate and adding a recording's first log time and associated trip's first midnight. Some error is introduced through this process as the data logger may start recording footage a couple of seconds later (not by a constant amount), and data gaps may go unnoticed and could accumulate a data offset as large as a few minutes before the end of a trip. While time stamps are theoretically retrievable from a watermark baked into the recordings at a 1s resolution, automated watermark extraction has not yet been performed. The third data category contains derived data available on a 10Hz table (road type, maximum permitted speed, gear setting). Time stamps are available as datetime and time in trip instead of midnight and logtime.

The various data sources were resampled through nearest neighbour selection to a common reference, which for the Tesla data was the estimated time in trip of the video-extracted automation status, while for the BMW time in trip according to the 10Hz table was used. NaN values were adopted whenever time differences spanned more than twice the source's sampling rate, when the source did not contain data for the desired variable at the matched timestamp or when the value was judged as unreliable (e.g. map matching error >50 m). Data was not included when automation status was unavailable or when not marked as to be used for this dataset.

Despite best efforts, a couple of faults still persist in the data as presented in this paper. The datetime stamps of the 10Hz table are calculated incorrectly from the midnight and logtime values. The 10Hz table also contains large sections of missing data where the same data is available in the raw tables. Speculations to the possible cause of this include: 1) only populating the 10Hz table for time stamps where GPS fix was available (GPS availability adopted the same variable name as the variable indicating if data was allowed to be analysed), 2) mismatched data association due to the incorrect time conversion, 3) 1 or 2 h offsets due to SQL server's default assumption of treating date times as system time instead of Unix time 4) overriding data from multiple data loggers by not ensuring unique trip and recording identifiers and addressing data by reference instead of by value. Because of these uncertainties, as little use of the 10Hz table is made as possible. For the BMW, raw automation status data is incomplete as indicated by an off-balance in data availability of ACC status and LK status. The suspected reason of this is a data re-structuring on the automation tables that took place at the time the dataset was exported to SWOV. Finally, many trip continuations appear to be missing for the BMW data as indicated by the sudden drop in data availability exactly 30 minutes into a trip (Figure 3.8). Possible causes are yet to be investigated.

4

Measuring Driver Perception: Combining Eye-tracking and Automated Road Scene Perception

Jork Stapel, Mounir El Hassnaoui, Riender Happee

This chapter has been published as: Jork Stapel, Mounir El Hassnaoui, Riender Happee, *Measuring Driver Perception: Combining Eye-tracking and Automated Road Scene Perception*, Human Factors, (2020).

Abstract

Objective. *To investigate how well gaze behavior can indicate driver awareness of individual road users when related to the vehicle's road scene perception.*
Background. *An appropriate method is required to identify how driver gaze reveals awareness of other road users.*
Method. *We developed a recognition-based method for labeling of driver situation awareness in a vehicle with road-scene perception and eye tracking. 13 drivers performed 91 left turns on complex urban intersections, and identified images of encountered road users among distractor images.*
Results. *Drivers fixated within 2° for 72.8% of relevant and 27.8% of irrelevant road users and were able to recognize 36.1% of the relevant and 19.4% of irrelevant road users one minute after leaving the intersection. Gaze behavior could predict road user relevance but not the outcome of the recognition task. Unexpectedly, 18% of road users observed beyond 10° were recognized.*
Conclusions. *Despite suboptimal psychometric properties leading to low recognition rates, our recognition task could identify awareness of individual road users during left turn maneuvers. Perception occurred at gaze angles well beyond 2° which means that fixation locations are insufficient for awareness monitoring.*
Application. *Findings can be used in driver attention and awareness modelling, and design of gaze-based driver support systems.*

PRÉCIS

To investigate how well gaze behavior can predict driver awareness of individual road users, participants drove 91 left turns on urban intersections in a vehicle equipped with road-scene perception and eye tracking, and identified images of encountered road users among distractor images.

4.1. INTRODUCTION

Perceptual errors contribute 76% of situation awareness (SA) errors (Jones and Endsley, 1996) and are among the most frequently reported causes for accidents at intersections, which represent 20% of European road accidents (European Road Safety Observatory, 2018). Vehicles are becoming more aware of their surroundings. Machine perception can locate road users through detection and classification systems (Liu et al., 2016; Kooij et al., 2014). It processes raw sensor data in a series of filters trained to extract features which collectively capture the concept of an object category. However, machine perception generally does not outperform human perception. Since the filters are trained from examples, they only function reliably in conditions similar to the training set. They also cannot comprehend what is seen, and only indicate if an object class occupies a particular region in the image. On the other hand, machine perception has superior attention in detection tasks. It can process the entire road scene without constraining to a region to attend, and does not suffer from vigilance decrement or biases from expectations. Machine perception can therefor support drivers in perceiving relevant road users through auditory and visual notifications. However, our senses receive more information than we can process with undivided and optimal fidelity, which we only overcome with a keen ability to be selective in what to attend. Augmentation of this process can only complement the driver effectively when it is equally selective, and becomes available well before the need is evidenced by a driver's inaction. To achieve this, driver support systems have to identify discrepancies between what is and what should be attended.

While considerable progress has been made in the development of systems to judge object relevance (Gao et al., 2019; Gary and James, 2019) or to redirect attention using audio (Ho and Spence, 2009), augmented reality (Kim et al., 2018) and peripheral displays (Yang et al., 2018), a key challenge lies in the decision when drivers need to be warned. Current systems rely on heuristics like "only alert when dangerous, rare or in conflict with common expectation", which generally limits operation to immediate hazards. Targeted support for developing hazards or non-critical lapses can only be achieved when driver awareness towards individual road users is monitored.

Eye tracking seems to be an ideal method to monitor what drivers have seen or overlooked, since people tend to fixate at what they inquire information from. De Winter et al. (2018) showed that glance behavior correlated better with supervision performance than the popular Situation Awareness Global Assessment Technique (SAGAT). Meghanathan et al. (2019) demonstrated that refixation patterns can discriminate encoding and memorization activity, and indicate change detection performance. However, fixation location does not always correspond to what is processed cognitively (Endsley, 1988; Rumar, 1990). Peripheral vision can suffice for lane keeping (Summala et al., 1996) and hazard detection (Huestegge and Böckler, 2016). Conversely we can fail to see things we fixate on (Mack and Rock, 1998), but it is yet unknown how frequently drivers miss other road users despite fixating upon them, or how to infer this from gaze behavior.

While aggregate metrics like distraction or fatigue have been inferred from vehicle-fixed regions of interest or direction independent measures like gaze variance (Rendon-Velez et al., 2016), gaze-based awareness classification of individual objects remains an open challenge. Attention prediction models like top-down saliency maps (Xia et al., 2019) or (N)SEEV (Wickens et al., 2007; Wickens, 2015) can compare current and nominal gaze behavior. When used to evaluate attention, the assumption is made that modelling what commonly *is* attended represents what *should* be attended. While this is reasonable for normal conditions, it may fail in error-prone or expectation-defying scenarios.

Hooey et al. (2011) evaluates situation awareness (SA) as a ratio between actual and optimal awareness among individual situational elements, weighted by their relevance. Aspects of saliency, expectancy and effort are not incorporated to predict likelihood of gaze, but to estimate difficulty of perception and comprehension. However, this approach often assumes a simple threshold of fixation eccentricity or duration to signify perception, and lacks quantitative calibration or validation (Fletcher and Zelinsky, 2009; Wickens et al., 2007). To understand how awareness can be inferred from gaze, large scale ground-truth labeling of SA is needed.

A variety of techniques exist to obtain such SA labeling (Stanton et al., 2013, Chapter 7; Nguyen et al., 2019). However, we believe that currently there is no suitable method for on-road, per-object awareness assessment. Physiological measures of SA lack construct specificity. One possible exception is EEG, which can track attention allocation to audio (Lu et al., 2018) and detect perception of hazards, conflicts or errors (Spüler and Niethammer, 2015; Wessel, 2012), but is not sufficiently discriminative to reliably detect single events. Self-rating and observer rating techniques are limited to aggregate measures rather than per-object assessments.

Freeze probe techniques like SAGAT measure object specific SA, but can only be used in simulators and measure recollection instead of awareness (de Winter et al., 2018). Recollection cannot probe unconscious/implicit awareness and suffers from inaccuracies like forgetting (Nisbett and Wilson, 1977), limiting it to partial scene probing (Gugerty, 1998). Real-time probes like SPAM (Durso and Dattel, 2004; Strybel et al., 2016) or verbal protocol methods (Salmon et al., 2014) circumvent these issues. However, the real-time communication is intrusive and limits probe rate. Furthermore, non-choreographed scenarios require that questions are generated real-time, as demonstrated by Sirkin et al. (2017) who automatically generated questions requiring simple yes/no and touch responses from the driver.

For this study, we build upon this idea of computer generated queries for unchoreographed on-road driving, and extend it to 1) enable the assessment of all relevant situational elements as opposed to sampling one at a time and 2) not distract the driver visually or cognitively while driving, so that it can be applied to complex maneuvers without overloading the driver. It is applicable to any driving scenario, but we apply it to left turns on urban intersections.

To prevent dangerous distraction, the probing task was performed after crossing the intersection and parking the car. This delay meant the driver had to memorize what transpired for longer compared to freeze probe methods, which may lead to memory decay. To minimize effects of decay, we use a visual recognition task instead of a recall task. Visual detail can be encoded quite effectively. Brady et al. (2011) reviews that natural scenes can be consolidated into memory within 100-500 ms while Lyu et al. (2019) and Choe et al. (2017) show that such encoding occurs incidentally without an attempt to memorize, which supports the idea that recognition can probe implicit as well as explicit awareness (Campodonico and Rediess, 1996). Working memory tasks have demonstrated that encoding fidelity reduces as demand increases, and that encoding multiple objects simultaneously is particularly difficult (Brady et al., 2011). Change blindness tests have demonstrated that changing vehicle presence, location and orientation is noticed, but subtle color changes are not (Koustanaï et al., 2012). However, Konkle et al. (2010b; 2010a) also demonstrate that a recognition task allows participants to identify scenes and objects among similar decoys with high accuracy (87% and up) after briefly observing 2500 images.

For this study, we designed a simple recognition task, where the driver has to identify images of encountered road users among distractor images. Successful recognition requires that the road user was perceived explicitly or implicitly, and thus provides an indicator for Endsley's (1995b) Level 1 situation awareness.

4.2. RESEARCH OBJECTIVE

We aimed to gain insight in drivers' natural viewing at intersections and how well SA can be predicted from gaze metrics relative to individual road users.
The main research questions addressed were:

1. Can a recognition task be used to asses per-object awareness?
2. Can SA be predicted from gaze metrics relative to individual road users?

(a) To which extent can gaze metrics predict object relevance and object recognition?

(b) To which extent are foveal and peripheral vision effective in the detection of other road users?

We developed a new method measuring SA in an urban on-road driving environment, and evaluated how well a variety of object-related gaze parameters can predict recognition after left-turn maneuvers. Driver gaze was related to object location, type and relevance for safety, using the road scene perception of our experimental vehicle. We then assessed the distribution of central and peripheral detections of safety relevant and irrelevant road users.

4.3. METHOD

We designed an experiment in which participants drove a Toyota Prius instrumented for automated driving and road scene perception (Ferranti et al., 2019) and eye tracking. The drivers (manually) performed left turns at multiple crossroads while the vehicle collected gaze behavior in relation to other road users. After each turn the driver parked the vehicle, and an object recognition task was performed to measure awareness of other road users encountered on or near the intersection. In subsequent sections, we detail the tracking of gaze and road users, the implementation of the recognition task and the experimental procedures.

TRACKING OF GAZE AND OTHER ROAD USERS

We used a 4-camera Smart Eye Pro dx 5.0 eye tracker (software version 8.2) running at 60 Hz with a gaze accuracy down to 0.5 degrees. The vehicle interior and setup are shown in Figure 4.1.

Figure 4.1: Interior of the vehicle showing the four eye-tracker cameras encircled in red, and the built-in display used for the post-drive recognition task.

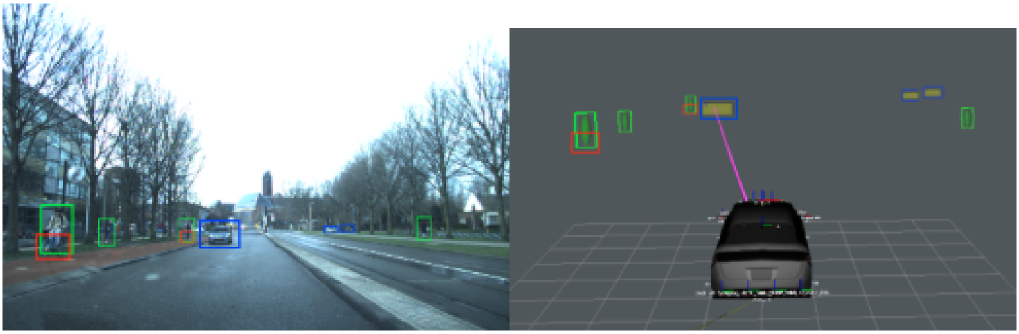


Figure 4.2: Left: road scene with highlighted detections. Right: 3D visualization of the detected objects. The magenta ray visualizes the driver's gaze.

The road was observed at 10 Hz using two forward facing IDS 2.3-megapixel cameras mounted near the top-center of the windshield and placed 22 cm apart for obtaining a dense stereo depth image over a visual angle of 62 degrees. Detection of other road users was performed using a single shot detector (Liu et al., 2016). Using the depth-image, detections were projected in 3D using the 15 percentile distance of all pixels inside the detection bounding box. Gaze analysis was limited to the horizontal component to reduce tracking artifacts caused by vehicle pitch motion at the cost of losing some specificity in the association of gaze angles to road users. After correcting for vehicle ego-motion, the road users were tracked in 3D space using a Kalman filter, and up-sampled to 60 Hz using linear interpolation in synchronization with the eye tracker. For each tracked object, the image with the largest bounding box was stored for use in the recognition task. Each image was made 20% larger than the bounding box to include some of the surrounding environment. It was then scaled to 200x200 pixels and normalized in brightness and contrast to reduce optical differences between real and dummy images. Information was integrated as visualized in Figure 4.2.

RECOGNITION TASK

After each intersection, the driver parked the vehicle, and the recognition task was performed on the vehicle centre display. The experimenter first prepared a selection among the road-user images which the vehicle had collected during the manoeuvre. The procedure was to avoid parked vehicles and blurry or partial images, and to select the clearest image/trajectory whenever multiple images were available for the same road user. The pre-selection GUI is shown in Figure 4.3. Each road user was represented with a colour image, its travelled path over the intersection and a summary of the road user properties (e.g. type, speed) and gaze parameters.

The selected images were then presented to the participant. Dummy images from an earlier session at that same intersection were added to discourage guessing. The participants were made aware that the task contained both road users they just encountered and dummy objects, and each dummy image was used once for each participant. The gaze details and vehicle data were not shown to the participant. The participant GUI is shown in Figure 4.4.

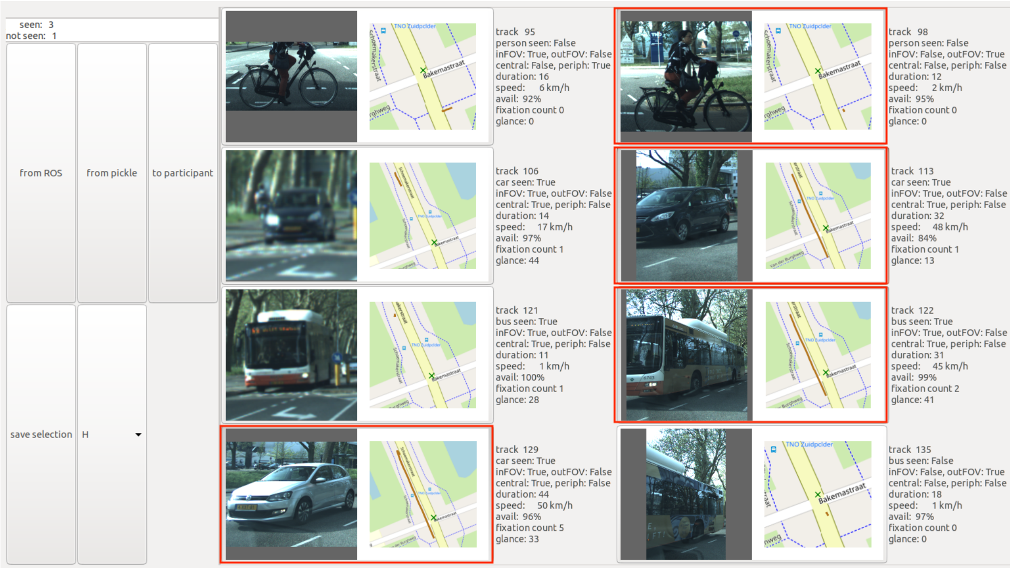


Figure 4.3: Example image of the pre-selection GUI in which the experimenter selected suitable object images to be presented to the participant. The GUI allowed scrolling up and down for additional images.

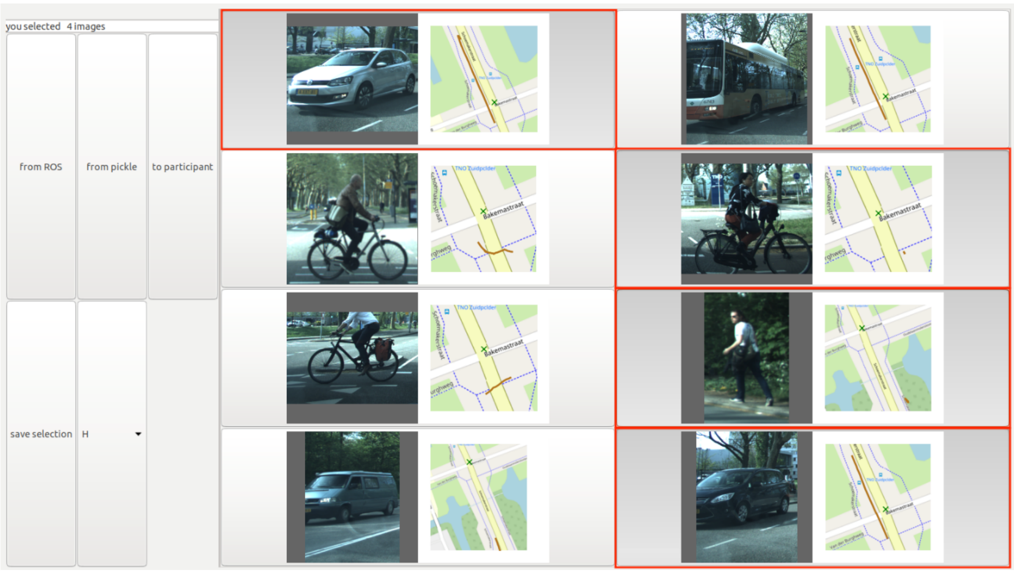


Figure 4.4: Example image of the GUI in which the participants selected images of the objects they recognized. 4 images are selected in this example. The GUI allowed scrolling up and down when more than 8 images were presented.

PROCEDURES

The criteria to participate in this study were: being a staff member or student of the department (for insurance reasons), having a driving license and having driven an automatic transmission at least once before. Fourteen male drivers participated, of which one was excluded from the analysis because the drive was not recorded correctly. The remaining 13 were aged between 24 and 57 years (M = 28.8, SD = 8.8). One had a license for 1-5 years, nine for 5-10 years and three for more than 10 years. Five participants drove less than once a month, four drove once a week, three drove 1-3 times a week and one participant drove every day. Four participants did not wear any visual aids, four wore glasses and four wore contact lenses. The research was approved by the Ethics Committee of the TU Delft. All participants read and signed an informed consent form prior to the experiment. They received a box of chocolates for their participation.

All participants were informed on the purpose of this study prior to participation and had a technical understanding of the used technology, but not of the recognition task or its implementation. Eye tracker calibration typically resulted in 1.2° accuracy and was repeated when accuracy exceeded 2° for at least one eye. They were navigated by the experimenter. Upon approaching each intersection, data recordings were started. The participants were asked to make a left turn and then safely stop or park the car at the first opportunity. They were asked (and asserted by the experimenter) to not look at the display while the recognition task was prepared. Once ready, the participants performed the recognition task without time constraint. The participants then returned to the main road and the procedure was repeated for all intersections. The participants returned to the starting location, where they completed a personal information questionnaire, rated the difficulty of the recognition task and indicated if they used the images, maps or both for their decisions.

The driven route is shown in Figure 4.5. Five intersections on the Schoemakerstraat in Delft were selected for their complexity, similarity and presence of traffic throughout the day. Three intersections were T-junctions that were passed once, and two were crossroads that were passed two times from opposite directions. The drivers had to give priority to oncoming traffic on the main road, to cyclists on the two-way bicycle path and pedestrians on the sidewalk. A typical intersection is shown in Figure 4.6. Maneuver 1 was an additional right turn used to practice the recognition task and was not analyzed.

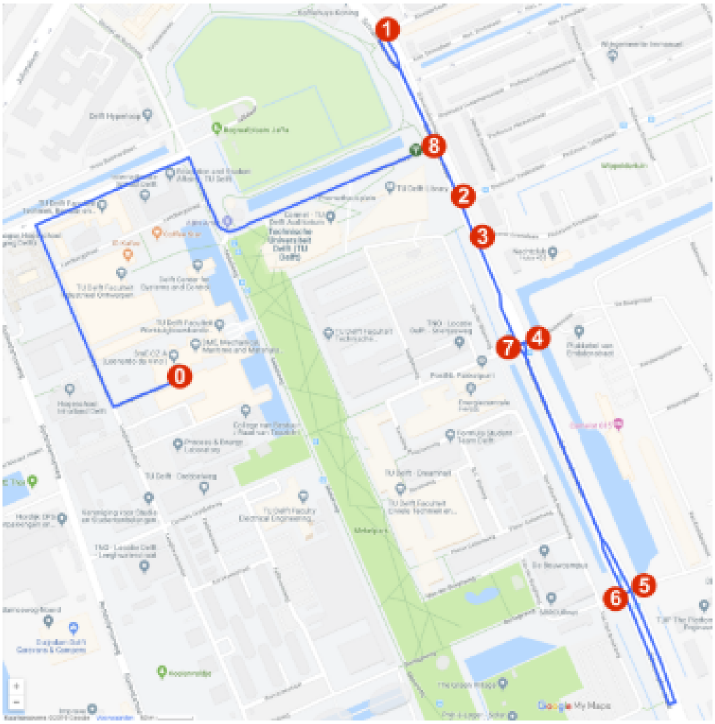


Figure 4.5: Map of the experiment driving route and the locations and order of the maneuvers.

Figure 4.6: One of the intersections. The green line corresponds to maneuver 6. (Google Street-View, 2019)

FILTERING AND MERGING OBJECTS

After the experiment, the collected data were filtered manually. Split or duplicate tracks of the same road user were merged. All road users were annotated as being relevant or irrelevant to the driving maneuver. The second author subjectively judged if a driver would want to monitor each road user for the purpose of driving at any time during the maneuver. A road user was considered relevant if the annotator felt that the participant had to give priority or should obtain priority at the intersection. Road users that left the intersection before the participant arrived at the intersection or were still well away when the participant left the intersection were regarded as irrelevant. Road users on the sidewalk or the bicycle lane on the right side of the intersection were annotated as irrelevant. Road users trying to enter or cross the main priority road before the participant passed were annotated as relevant. An overview of the possible road users encountered and how they were annotated is shown in Figure 4.7.

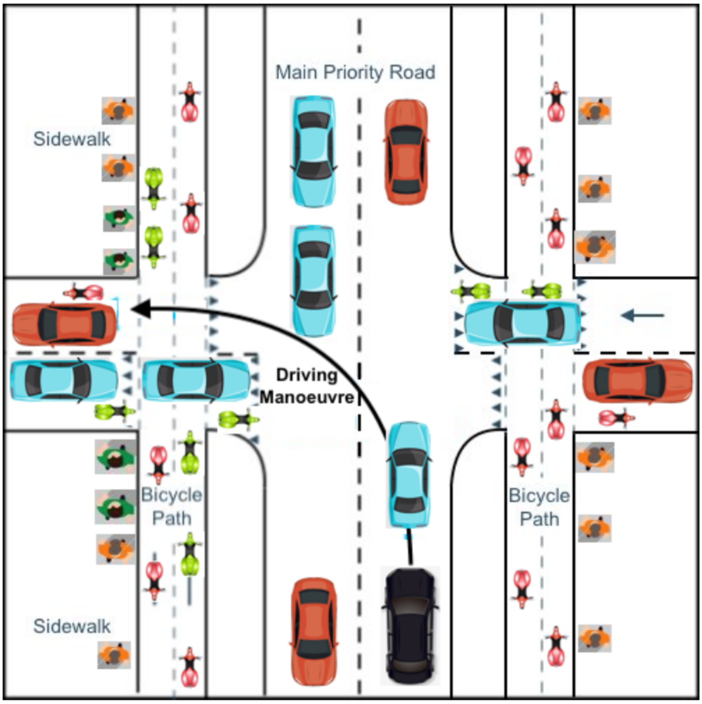


Figure 4.7: Schematic visualization of how objects were annotated according to their position and movement direction: relevant road users (blue/green) and irrelevant road users (red) and the car driven by the participant (black).

GAZE METRICS

The following parameters were analyzed

Gaze eccentricity: the angle between the direction of the driver's gaze and the vector from the gaze origin to the center of a road user (width of the road user is ignored). Only the horizontal component is used in this study.

Minimum gaze angle: the smallest gaze eccentricity towards a road user throughout the period this road user was tracked by the vehicle. Saccades (angular rates beyond 35 deg./s) are ignored.

Total glance duration within visual field regions: the summed duration of all fixations occurring while the gaze eccentricity falls within one of the following regions (Duchowski, 2007): foveal view (<2°) where highest visual acuity is obtained, near-foveal view (2 - 5°) in which objects are commonly recognizable, central view (5 - 10°) up to which acuity and color sensitivity degrade linearly, near-peripheral view (10 - 30°) and far-peripheral view (>30°)

Huestegge and Böckler (2016) compared saccade behavior during the detection of critical and moderate hazards in static scenes and found that more critical hazards are detected earlier, at larger peripheral angles and with shorter fixation durations preceding the first saccade to these hazards. We therefore evaluated the related metrics:

First saccade angle: the visual angle between start and end of the first saccade that lands within 2° of the object.

First saccade time: the time over which the object has been tracked by the vehicle before a first saccade lands within 2° of the object. Any saccade landing on the object before it was detected by the vehicle is not observable and thus ignored.

Duration preceding fixation: the duration of the fixation that preceded the first saccade landing within 2° of the object.

We also observed the following parameters that are often considered to assess situation awareness. They were excluded from regression analysis because they are structurally correlated to the total glance duration within 2° . Instead, simple effects are reported.

Number of fixations: the number of fixations occurring while the object is within 2° of the gaze vector. It is equivalent to the number of saccades

Mean glance duration: total glance duration within 2° divided by the number of fixations.

BINARY LOGISTIC REGRESSION

Binary logistic regression was performed to test if the gaze parameters can predict the participants' selections in the recognition task. We also tested if gaze parameters can predict object relevance. To account for subject dependencies, both models use participant as a random variable for the intercept.

We had to address missing values for saccade-related variables, which are defined only when they land within 2 degrees of an object. List-wise elimination is not desired since we want a prediction even for objects that were not glanced upon directly. Instead we adopted Cohens' dummy-variable adjustment (Cohen and Cohen, 1983). This approach is not generally recommended, as it may induce bias from conditional inclusion (Allison, 2009). In our case however, such a bias is not a concern as the missing values are a structural property of the model. The model structure thus becomes:

$$Y = b_{0\,\mathrm{j}} + b_1 X_1 + Z\,(b_2 X_2) + e$$

Here b_{0j} represents the intercept which is allowed to vary among participants, X_1 are total glance durations for the 5 eccentricity ranges, X_2 are saccade related variables and Z the dummy variable where $Z=1$ when saccades are available and 0 otherwise.

RESULTS

In the recognition task, participants had to select images of road users they just encountered. A total of 91 intersection crossings were collected and 1824 images were presented to the 13 participants in total. On average, there were 8.2 images of real objects and 11.8 dummy images per intersection per participant. The number of images presented to the driver varied with a standard deviation of 6 and ranged between 5 and 34. It took 30 seconds on average to park the car after leaving the intersection, and another 30 seconds for the experimenter to prepare the recognition task. Participants took approximately 80 seconds to select images. The questionnaire showed that the participants rated the difficulty of the recognition task as 8.6 on a scale of 1 to 10, with 1 being really easy and 10 being really difficult. Due to dropped messages in the recordings, 7% of the gaze data could not be re-associated to the recognition task images and were omitted from the gaze-related analysis.

SELECTION OF IMAGES

Table 1 shows how often participants selected images of real and dummy objects and provides an indication of response bias and sensitivity. The odds of selecting an image was 5.7 times higher for real compared to dummy images ($CI_{95\%}$= 4.3, 7.6). Only 29.1% of real images were selected. Since no unsafe driver behavior was noted, the remaining 70.9% do not necessarily represent overlooked road users. Hence, recognition rates reported in this study must under-estimate the actual SA and our recognition task can thus not fully address research question 2.

The 93.3% not selected dummy objects suggest that the participants adopted a select-only-when-certain philosophy. While the 72 selected dummy images could result from guessing, some may have been confused with real objects. 13.7% of these dummy images shared a close resemblance to a real image, similar to Figure 4.8. 24.7% had an approximate resemblance to a real image of same type, approximately sharing color and/or shape. Jointly this suggests that selected images indeed represent perceived road users.

Relevant (real) road users were recognized more often (36.1%) than irrelevant ones (19.4%). This interaction effect is significant ($\chi^2(1)$= 26.06, p=<0.001, φ_c=0.186) with an odds ratio of 2.3 ($CI_{95\%}$ = 1.7, 3.3).

Table 4.1: Contingency table of the selected images of real objects and dummy objects, as well as relevant and irrelevant objects.

	Selected		Not selected	
Real images	218	29.1%	532	70.9%
Relevant objects	144	36.1%	255	63.9%
Irrelevant objects	74	19.4%	307	80.6%
Dummy images	72	6.7%	1002	93.3%

Figure 4.8: Left: an example of a selected dummy image; a silver Volkswagen. Right: a not selected real image actually encountered; a silver Toyota.

MINIMUM GAZE ANGLE

Table 2 shows the number of real road users divided into the different object classes and minimum gaze angles. 27.2% of the relevant and 72.2% of the irrelevant road users were never fixated upon within 2°. Similarly, 40.1% of the recognized and 53.7% of the not recognized road users were never fixated upon within 2°. These values are surprisingly large and indicate that a considerable number of objects were perceived without ever receiving a direct fixation. Cars had the lowest recognition rate (26.2%) despite being the most common. Busses are recognized the most (61.5%), followed by motorcycles (44.4%) and pedestrians (39.0%). When only considering relevant road users, pedestrians were recognized the most (64.3%). Minimum gaze angle interacted significantly with recognition ($\chi^2(4)$= 16.07, p=.003, φ_c=0.151), suggesting that higher eccentricity leads to poorer recognition. Minimum gaze angle also interacted with relevance ($\chi^2(4)$= 151.35, p=<.001, φ_c=0.463), suggesting that relevant objects are monitored more closely compared to irrelevant ones. Road user type also interacted significantly with recognition (Fisher's exact test = 15.13, p=.020, φ_c=0.151) and relevance (Fisher's exact test = 51.58, p<.001, φ_c=0.266).

Figure 4.9 shows the distribution of gaze angles over the duration that an object was tracked by the car; averaged over all tracked objects. Driver gaze dwelled closer to relevant compared to irrelevant road users. A similar effect is not as clear between selected and not selected road users.

4

Table 4.2: Number of real road users observed at various minimum gaze angles, for all objects (left) and those selected during the recognition task (right). The “other” category comprises one dog and two excavators.

	All objects (minimum gaze angle)						Recognized objects (minimum gaze angle)					
	N	<2°	2-5°	5-10°	10-30°	>30°	N	<2°	2-5°	5-10°	10-30°	>30°
Car	409	257	62	39	40	11	107	76	14	9	6	2
Relevant	241	191	23	15	9	3	81	63	9	6	2	1
Irrelevant	168	66	39	24	31	8	26	13	5	3	4	1
Bicycle	184	67	24	27	57	9	54	31	8	5	8	2
Relevant	83	48	11	9	13	2	39	29	5	3	2	0
Irrelevant	101	19	13	18	44	7	15	2	3	2	6	2
Pedestrian	77	17	12	10	32	6	30	10	7	3	9	1
Relevant	14	7	2	2	3	0	9	5	2	2	0	0
Irrelevant	63	10	10	8	29	6	21	5	5	1	9	1
Bus	13	7	2	2	2	0	8	4	2	2	0	0
Relevant	5	5	0	0	0	0	3	3	0	0	0	0
Irrelevant	8	2	2	2	2	0	5	1	2	2	0	0
Truck	11	3	2	1	4	1	3	1	0	0	2	0
Relevant	6	3	1	0	2	0	2	1	0	0	1	0
Irrelevant	5	0	1	1	2	1	1	0	0	0	1	0
Motor	9	4	1	3	1	0	4	2	1	1	0	0
Relevant	4	3	0	1	0	0	2	2	0	0	0	0
Irrelevant	5	1	1	2	1	0	2	0	1	1	0	0
Other	3	0	0	1	2	0	1	0	0	1	0	0
Relevant	0	0	0	0	0	0	0	0	0	0	0	0
Irrelevant	3	0	0	1	2	0	1	0	0	1	0	0
Total	706	355	103	83	138	27	207	124	32	21	25	5
Relevant	353	257	37	27	27	5	136	103	16	11	5	1
Irrelevant	353	98	66	56	111	22	71	21	16	10	20	4

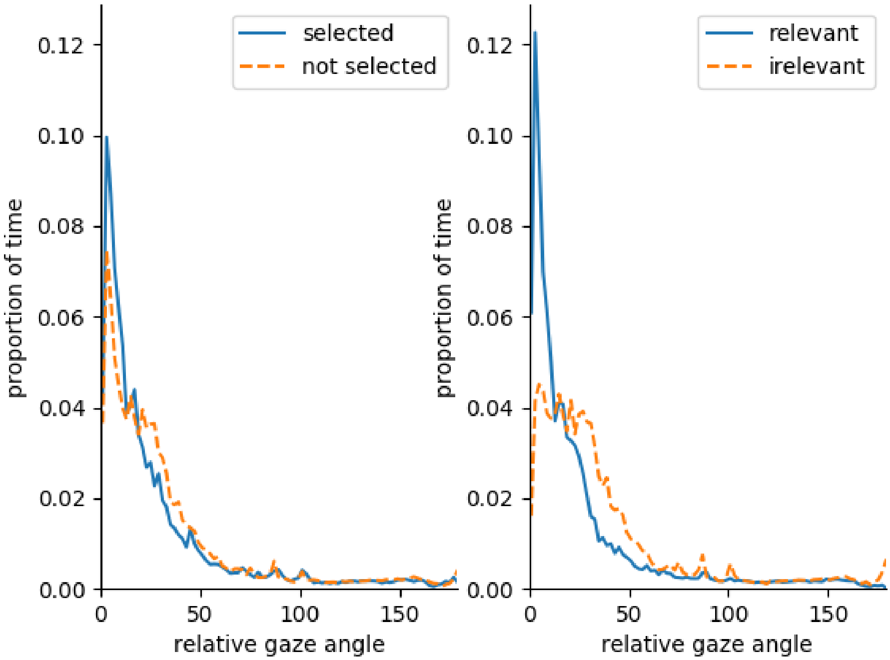


Figure 4.9: Distribution of relative gaze angle for (not) selected (left) and (ir)relevant (right) road users over the period they were detected by the vehicle cameras.

FIXATION PARAMETERS

Table 3 compares fixation parameters for relevant vs irrelevant and selected vs not selected road users. Figure 4.10 and Figure 4.11 illustrate the distribution shapes for a selection of gaze parameters comparing relevance and selection in the recognition task. All are right-tailed. The Kolmogorov-Smirnov tests in Figure 4.11 represent a non-parametric statistic of similarity between the paired distributions. Only total glance duration with fixation angle <2° yielded a significantly different distribution between relevant and irrelevant road users. To avoid dependencies among participants and reduce non-normality of the distribution of the residuals, we used participant-averaged paired t-tests. Relevant road users received 1.14 more fixations compared to irrelevant road users. Selected road users received 0.41 more fixations compared to not selected road users. Mean fixation duration did not differ significantly in either case.

Table 4.3: Fixation parameter mean (μ) and standard deviation (σ) as function of relevance (top) and being selected (bottom).

	Relevant		Irrelevant			
	μ	σ	μ	σ	T(12)	*p*
Number of fixations <2°	1.59	0.51	0.45	0.26	9.653	<.001
Total fixation duration <2° (ms)	955	309	237	127	9.318	<.001
Mean fixation duration (ms)	658	264	685	676	−0.134	.895
	Selected		**Not selected**			
	μ	σ	μ	σ	T(12)	*p*
Number of fixations <2°	1.34	0.51	0.93	0.33	2.341	.037
Total fixation duration <2° (ms)	833	519	536	197	1.944	.076
Mean fixation duration (ms)	633	272	640	257	−0.113	.912

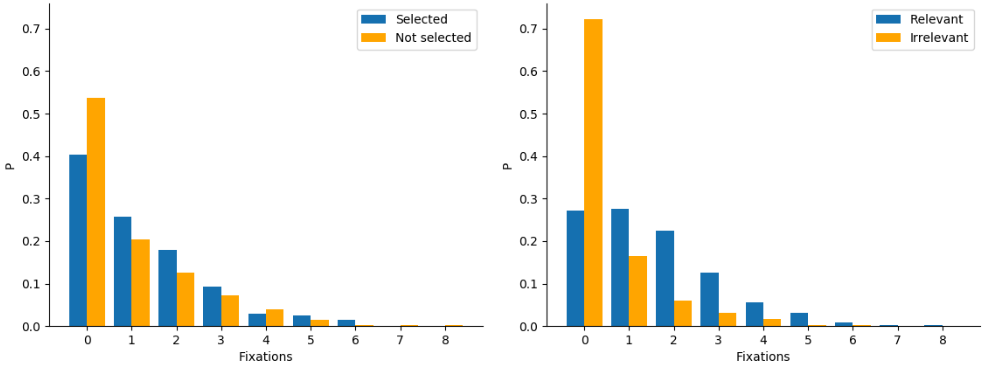


Figure 4.10: Distribution of fixations (<2°) per road user, comparing selection in the recognition task (left) and relevance (right).

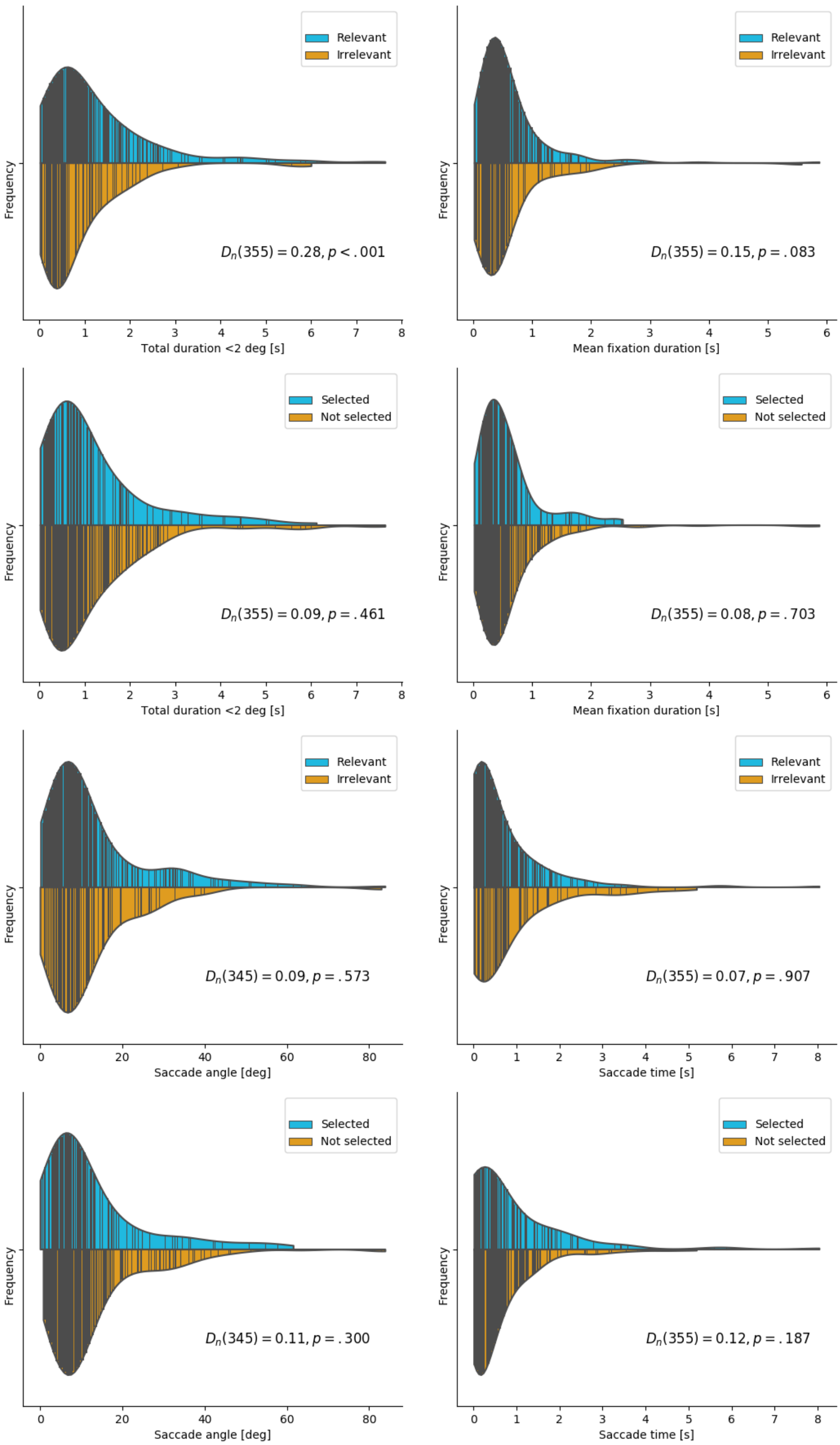


Figure 4.11: Distributions and Kolmogorov-Smirnov tests for a selection of gaze parameters, comparing relevance and selection in the recognition task.

Binary logistic regression

Table 4 shows the classification performance for the recognition and the relevance models. Model accuracy was compared to intercept models, which obtained an accuracy of 70.7% by simply predicting that no objects were selected, and an accuracy of 50.0% by predicting that all objects were relevant. Both models differed significantly from their intercept models (Recognition: $\chi^2(1)$=37.64, p<.001, φ_c=0.231. Relevance: $\chi^2(1)$=152.49, p<.001, φ_c=0.465), where accuracy increased by only 2.12% to 72.80% for the recognition model and by a more substantial 23.09% to 73.09% for the relevance model.

Table 4.4: Classification performance of the logistic regression models.

	Intercept models				Parameterized models			
	Relevance		Recognition		Relevance		Recognition	
Predicted	True	False	True	False	True	False	True	False
Observed								
True	353	0	0	207	238	115	44	163
False	353	0	0	499	75	278	29	470

Table 4.5: Parameters of the logistic regression models.

	Relevant					Recognized				
	Exp(b)	t	p	5%CI	95%CI	Exp(b)	t	p	5%CI	95%CI
Intercept	0.390	−3.629	<.001	0.234	0.649	0.262	−4.274	<.001	0.141	0.484
Duration <2° [s]	5.452	3.024	.003	1.813	16.398	1.424	1.591	.112	0.921	2.204
Duration 2-5° [s]	2.658	3.273	.001	1.479	4.778	0.956	−0.157	.875	0.544	1.679
Duration 5-10° [s]	2.541	4.188	<.001	1.641	3.934	1.995	3.153	.002	1.298	3.067
Duration 10-30° [s]	1.094	0.741	.459	0.862	1.390	0.946	−0.402	.688	0.720	1.242
Duration >30° [s]	0.693	−1.574	.116	0.439	1.095	0.929	−0.444	.657	0.670	1.288
1st Saccade angle [°]	1.049	2.756	.006	1.014	1.085	1.001	0.132	.895	0.982	1.021
1st Saccade time [s]	0.901	−0.613	.540	0.646	1.257	1.243	1.334	.183	0.903	1.711
Preceding fixation [s]	1.087	0.345	.730	0.677	1.744	1.361	1.217	.224	0.828	2.239

Table 5 provides the exponentials and significance of the model parameters. The odds for a road user being relevant increases significantly with gaze duration in relative gaze angle ranges of <2°, between 2-5° and between 5-10°. Gazes at larger angles do not seem to discriminate between relevant or irrelevant road users. The odds for a relevant road user also increases slightly when the first saccade within <2° has a larger angle. Timing of the first saccade or the duration of its preceding fixation do not help to discriminate relevance of road users.

The odds of recognizing a road user increases significantly only when the relative gaze angle spends more time between 5-10° from the road user, and a similar effect for gaze <2° does not reach significance (p=.11). Since this model used 9 parameters to only achieve a 2% accuracy improvement over the intercept model, the relevance of these results is limited.

DISCUSSION

This study set out with two objectives: to evaluate if the developed recognition task can provide useful labeling of per-object situation awareness and to evaluate whether awareness of other road users can be predicted from gaze behavior in relation to these objects.

SUITABILITY OF THE RECOGNITION TASK

Relevant road users were recognized more often than irrelevant road users, which is in line with Moore and Gugerty (2010). However, drivers recognized only 29.1% of all road users, 36.1% of the relevant road users and 40.0% of relevant road users that were fixated <2°. These unexpected low recognition rates mean that the current implementation of the recognition task is only partially successful in labeling situation awareness. Below we analyze the limited recognition rate, and provide suggestions to adapt the task to enhance recognition.

While the vehicle processed all video and gaze measurements in real time, 60 seconds elapsed between finishing the maneuver and performing the recognition task. This time was needed for the participant to stop the vehicle and for the experimenters to select images for the recognition task. This delay may have contributed to the low recognition rate. Humans are normally poor in remembering details of past events with a rapid decay of information in working memory which is limited to around 30 seconds (Nisbett and Wilson, 1977). In contrast, Endsley suggests that SAGAT-like techniques do not suffer much from memory decay up to 3 minutes, provided that the participant is experienced (Endsley, 2019; 1995a). Delays below 30 seconds should be feasible if the selection of suitable images is automated, and a location is reserved after each intersection for faster parking.

Secondly, it is possible that our image representation differed too much from how situations are encoded by experienced drivers. Performing a left-turn on a busy priority road is relatively demanding. Working memory makes trade-offs between the quantity of stored items and their fidelity. The more road users we encounter, the fewer details about them we can store, and task-irrelevant features are the first to be dropped (Brady et al., 2011). Our images contained little task-relevant context. Although the maps provided some spatial context, all participants reported to primarily base their decisions on the images. A possible improvement would be to show more environment in the images, and project the travelled paths into the images instead of the separate map.

SUITABILITY OF GAZE BEHAVIOUR

We parameterized gaze behavior relative to nearby road users. Such gaze parameters may be useful in driver attention and awareness modelling and driver support system design.

Gaze behavior could predict object relevance with an accuracy of 73%, where relevant objects were more often fixated <2°, with larger 1^{st} saccade angles, and with a higher gaze duration up to 10° eccentricity. This illustrates that relevant road users are kept more within the useful field of view compared to irrelevant road users. Mean 1^{st} saccade amplitude was 12.6°, with a strongly skewed distribution well into the 30° region. This suggests that peripheral vision was effectively used to direct gaze to relevant road users.

While the first saccade angle contributed significantly to the relevance model, timing of the first saccade and duration of the preceding fixation did not. This difference with the findings of Huestegge and Böckler (2016) could mean that the usefulness of saccade parameters is limited to hazards. Saccade related parameters may also become more useful when studied in relation to events like changes in the road users' behavior instead of their first appearance in the driving scene.

Gaze behavior was not very effective in predicting outcomes of the recognition task. One explanation is that the forgetting aspect could not be captured by our model. We expect that improved methods and simpler conditions can enhance recognition rate and further clarify the relation between recognition and gaze. Meanwhile, the recognition task did provide useful insights. 18% of the road users that never entered the useful field of view (<10°) were still selected in the recognition task, highlighting the importance of peripheral vision (Wolfe et al., 2017). Hence, we strongly recommend that perception models incorporate more than fixation location in their parameterization. Our findings may also provide guidance for designing a system alerting drivers towards other road users which may be unseen. From Table 2 we estimate how frequently such a support system might alert drivers. Our dataset includes 8.2 road users per intersection of which 4.4 are relevant to the maneuver. When a gaze-aware system alerts to every peripheral (>10°) road user, the driver receives 1.8 alerts per intersection. Since drivers recognized 18% of the peripherally observed road users, they would have been aware of at least 0.33 alerts beforehand. When only responding to relevant peripheral road users, the driver receives only 0.35 alerts per intersection and would be aware of only 0.07 alerts beforehand.

LIMITATIONS AND FUTURE RECOMMENDATIONS

The recognition task can be improved to obtain complete rather than partial labeling of per-object situation awareness in complex unstructured maneuvers. The main limitations - delay before start of the test and task visualization - are likely to be overcome. Better object detection and especially more robust tracking could circumvent the need for manual preselection of candidate images, and thus reduce the delay between actual encounters and the recognition task. Further improvements may be achieved by reducing the number of test images as recommended by Gugerty (1998) or by associating gazes more selectively to a single road user, for instance through including the vertical gaze component, or with a Dynamic Markov random field model (Jiang et al., 2018) or Bayesian likelihood model (Schwehr and Willert, 2017).

After such improvements, the potential of the recognition task as a variant of freeze-probe methods can be explored. Benefits may emerge in simpler maneuvers or in the better controlled simulator environment. Improvements of the recognition task's visualization may better suit the driver's encoding of situational information. To improve retrieval for recognition, road users could be depicted in the road scene (Hollingworth, 2006), and road-user motion could be encoded using video, animation or multiple images.

Finally, to more closely examine time-critical attention allocation, it may be interesting to study gaze behavior relative to road user actions (such as a change in travelled path or speed) in addition to the aggregate road user parameters used in this study.

KEY POINTS

- Gaze relative to surrounding traffic was compared to a recognition task during on-road left-turn maneuvers
- Gaze behavior could predict object relevance where relevant road users were kept longer in the useful field of view
- Drivers recognized 18% of the peripherally observed road users, which suggests that perception models should consider more than foveated vision.
- Driver feedback can become more selective when driver awareness of individual road users is monitored.

Acknowledgment Funding: This work was supported by the NWO-TTW Foundation, the Netherlands, under the project “From Individual Automated Vehicles to Cooperative Traffic Management - Predicting the benefits of automated driving through on-road human behavior assessment and traffic flow models (IAVTRM)" -STW#13712.

4

4.4. REFERENCES

Allison, P.D. (2009) ‘Missing data’, in Millsap, R.E. and Maydeu-Olivares, A. (Eds.), *The SAGE handbook of quantitative methods in psychology*, SAGE publications

Brady, T.F., Konkle, T. and Alvarez, G.A. (2011) ‘A review of visual memory capacity: Beyond individual items and toward structured representations’, *Journal of vision*, Vol. 11, No. 5, pp.1–49

Campodonico, J.R. and Rediess, S. (1996) ‘Dissociation of implicit and explicit knowledge in a case of psychogenic retrograde amnesia’, *Journal of the International Neuropsychological Society*, Vol. 2, pp.146–158

Choe, K.W., Kardan, O., Kotabe, H.P., Henderson, J.M. and Berman, M.G. (2017) ‘To search or to like: Mapping fixations to differentiate two forms of incidental scene memory’, *Journal of vision*, Vol. 17, No. 12, p.8

Cohen, J. and Cohen, P. (1983) *Applied multiple regression/correlation analysis for the behavioral sciences*, 2^{nd} ed., L. Erlbaum Associates, Hillsdale, N.J

de Winter, J.C.F., Cabrall, C.D.D., Eisma, Y.B., Hancock, P.A. and Stanton, N.A. (2018) ‘Situation awareness based on eye movements in relation to the task environment’, *Cognition Technology & Work*, Vol. 34, p.117

Duchowski, A.T. (2007) *Eye tracking methodology. Theory and practice*, 2^{nd} ed., Springer, London http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10202868

Durso, F.T. and Dattel, A.R. (2004) ‘SPAM: The real-time assessment of SA’, in Banbury, S. and Tremblay, S. (Eds.), *A cognitive approach to situation awareness:. Theory and application*, Ashgate Publisging Limited, pp.137–154

Endsley, M.R. (2019) ‘A systematic review and meta-analysis of direct objective measures of situation awareness: A comparison of SAGAT and SPAM’, *Human factors*, 18720819875376

Endsley, M.R. (1995a) ‘Measurement of situation awareness in dynamic systems’, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 37, No. 1, pp.65–84

Endsley, M.R. (1995b) 'Toward a theory of situation awareness in dynamic systems', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 37, No. 1, pp.32–64

Endsley, M.R. (1988) 'Situation awareness global assessment technique (SAGAT)', *Proceedings of the National Aerospace and Electronics Conference*, pp.789–795

European Road Safety Observatory (2018) *Traffic safety basic facts on junctions*

Ferranti, L., Brito, B., Pool, E., Zheng, Y., Ensing, R.M., Happee, R., Shyrokau, B., Kooij, J.F.P., Alonso-Mora, J. and Gavrila, D.M. (2019) 'SafeVRU: A research platform for the interaction of self-driving vehicles with vulnerable road users', *IEEE Intelligent Vehicles Symposium*, pp.1660–1666

Fletcher, L. and Zelinsky, A. (2009) 'Driver inattention detection based on eye gaze-road event correlation', *The International Journal of Robotics Research*, Vol. 28, No. 6, pp.774–801

Gao, M., Tawari, A. and Martin, S. (2019) *Goal-oriented object importance estimation in on-road driving videos*. http://arxiv.org/pdf/1905.02848v1

Gary, C. and James, C. (2019) 'Traffic risk assessment: A two-stream approach using dynamic-attention', *IEEE Conference on Computer and Robot Vision*, Vol. 19 [online] https://doi.org/10.1109/CRV.2019.00030

Google Street-View (2019). https://www.google.com/maps/@51.998422,4.3795668,21z

Gugerty, L.J. (1998) 'Evidence from a partial report task for forgetting in dynamic spatial memory', *Human factors*, Vol. 40, No. 3, pp.498–508

Ho, C. and Spence, C. (2009) 'Using peripersonal warning signals to orient a driver's gaze', *Human factors*, Vol. 51, No. 4, pp.539–556

Hollingworth, A. (2006) 'Scene and position specificity in visual memory for objects', *Journal of experimental psychology. Learning, memory, and cognition*, Vol. 32, No. 1, pp.58–69

Hooey, B.L., Gore, B.F., Wickens, C.D., Scott-Nash, S., Socash, C., Salud, E. and Foyle, D.C. (2011) 'Modeling pilot situation awareness', *Human modelling in assisted transportation*, pp.207–213

Huestegge, L. and Böckler, A. (2016) 'Out of the corner of the driver's eye: Peripheral processing of hazards in static traffic scenes', *Journal of vision*, Vol. 16, No. 2, p.11

Jiang, Y.-S., Warnell, G. and Stone, P. (2018) 'DIPD: Gaze-based intention inference in dynamic environments', *AAAI Conference on Artificial Intelligence*, Vol. 32

Jones, D.G. and Endsley, M. (1996) 'Sources of situation awareness errors in aviation', *aviation, space, and environmental medicine*, Vol. 67, No. 6

Kim, H., Gabbard, J.L., Anon, A.M. and Misu, T. (2018) 'Driver behavior and performance with augmented reality pedestrian collision warning: An outdoor user study', *IEEE Transactions on Visualization and Computer Graphics*, pp.1–10

Konkle, T., Brady, T.F., Alvarez, G.A. and Oliva, A. (2010a) 'Conceptual distinctiveness supports detailed visual long-term memory for real-world objects', *Journal of experimental psychology. General*, Vol. 139, No. 3, pp.558–578

Konkle, T., Brady, T.F., Alvarez, G.A. and Oliva, A. (2010b) 'Scene memory is more detailed than you think: The role of categories in visual long-term memory', *Psychological science*, Vol. 21, No. 11, pp.1551–1556

Kooij, J.F.P., Schneider, N., Flohr, F. and Gavrila, D.M. (2014) 'Context-based pedestrian path prediction', *European conference on computer vision*, pp.618–633

Koustanaï, A., van Elslande, P. and Bastien, C. (2012) 'Use of change blindness to measure different abilities to detect relevant changes in natural driving scenes', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 15, No. 3, pp.233–242

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A.C. (2016) 'SSD: Single shot multibox detector', *European conference on computer vision* [online] https://github.com/weiliu89/caffe/tree/ssd

Lu, Y., Wang, M., Zhang, Q. and Han, Y. (2018) 'Identification of auditory object-specific attention from single-trial electroencephalogram signals via entropy measures and machine learning', *Entropy*, Vol. 20, No. 386

Lyu, M., Choe, K.W., Kardan, O., Kotabe, H.P., Henderson, J.M. and Berman, M.G. (2019) *Scenes that produce more consistent fixation maps are more memorable*

Mack, A. and Rock, I. (1998) *Inattentional blindness. An overview by Arien Mack & Irvin Rock*, MIT press, Cambridge

Meghanathan, R.N., Nikolaev, A.R. and van Leeuwen, C. (2019) 'Refixation patterns reveal memory-encoding strategies in free viewing', *Attention, perception & psychophysics*, Vol. 81, No. 7, pp.2499–2516

Moore, K. and Gugerty, L. (2010) 'Development of a novel measure of situation awareness: The case for eye movement analysis', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 54, pp.1650–1654

Nguyen, T., Lim, C.P., Nguyen, N.D., Gordon-Brown, L. and Nahavandi, S. (2019) 'A review of situation awareness assessment approaches in aviation environments', *IEEE Systems Journal*, Vol. 13, No. 3 [online] 10.1109/JSYST.2019.2918283

Nisbett, R.E. and Wilson, T.D. (1977) 'Telling more than we can know: Verbal reports on mental processes', *Psychological Review*, Vol. 84, No. 3, pp.231–259

Rendon-Velez, E., van Leeuwen, P.M., Happee, R., Horváth, I., van der Vegte, W.F. and de Winter, J.C.F. (2016) 'The effects of time pressure on driver performance and physiological activity: A driving simulator study', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 41, No. 41, pp.150–169

Rumar, K. (1990) 'The basic driver error: late detection', *Ergonomics*, Vol. 33, 10-11, pp.1281–1290

Salmon, P.M., Lenne, M.G., Walker, G.H., Stanton, N.A. and Filtness, A. (2014) 'Exploring schema-driven differences in situation awareness between road users: An on-road study of driver, cyclist and motorcyclist situation awareness', *Ergonomics*, Vol. 57, No. 2, pp.191–209

Schwehr, J. and Willert, V. (2017) 'Driver's gaze prediction in dynamic automotive scenes', *IEEE International Conference on Intelligent Transportation Systems*, Vol. 20, pp.1–8

Sirkin, D., Martelaro, N., Johns, M. and Ju, W. (2017) 'Toward measurement of situation awareness in autonomous vehicles', *CHI Conference on Human Factors in Computing Systems*, pp.405–415

Spüler, M. and Niethammer, C. (2015) 'Error-related potentials during continuous feedback: Using EEG to detect errors of different type and severity', *Frontiers in human neuroscience*, Vol. 9, No. 155

Stanton, N.A., Salmon, P.M., Rafferty, L.A., Walker, G.H., Baber, C. and Jenkins, D.P. (2013) *Human factors methods. Practical guide for engineering and design*, 2nd ed., Ashgate Publishing Limited, Wey Court East, Union Road, Farnham, Surrey, GU9 7PT England

Strybel, T.Z., Chiappe, D., Vu, K.-P.L., Miramontes, A., Battiste, H. and Battiste, V. (2016) 'A comparison of methods for assessing situation awareness in current day and future air traffic management operations: Graphics-based vs text-based online probe systems', *International Federation of Automatic Control*, Vol. 49, No. 19, pp.31–35

Summala, H., Nieminen, T. and Punto, M. (1996) 'Maintaining lane position with peripheral vision during in-vehicle tasks', *Human Factors and Ergonomics Society*, Vol. 38, No. 3, pp.442–451

Wessel, J.R. (2012) 'Error awareness and the error-related negativity: Evaluating the first decade of evidence', *Frontiers in human neuroscience*, Vol. 6, No. 88, pp.1–16

Wickens, C.D. (2015) 'Noticing events in the visual workplace: The SEEV and NSEEV models', in Hoffman, R.R. et al. (Eds.), *The cambridge handbook of applied perception research*, Cambridge University Press, New York, pp.749–768

Wickens, C.D., McCarley, J.S., Alexander, A.L., Thomas, L.C., Ambinder, M. and Zheng, S. (2007) 'Attention-situation awareness (A-SA) model of pilot error', in , *Human Performance Modeling in Aviation*, pp.213–239

Wolfe, B., Dobres, J., Rosenholtz, R. and Reimer, B. (2017) 'More than the useful field: Considering peripheral vision in driving', *Applied ergonomics*, Vol. 65, pp.316–325

Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K. and Whitney, D. (2019) 'Predicting driver attention in critical situations', *Computer Vision-ACCV*, pp.658–674

Yang, Y., Karakaya, B., Dominioni, G.C., Kawabe, K. and Bengler, K. (2018) 'An HMI concept to improve driver's visual behavior and situation awareness in automated vehicle', *International Conference on Intelligent Transportation Systems (ITSC)*, Vol. 21, pp.650–655

5

Driver and Pedestrian Joint Awareness for Collision Risk Prediction.

Markus Roth, Jork Stapel, Riender Happee, Dariu M. Gavrila

This chapter has been submitted as: Markus Roth, Jork Stapel, Riender Happee, Dariu M. Gavrila, *Driver and Pedestrian Mutual Awareness for Path Prediction and Collision Risk Estimation.*

Abstract

We present a novel method for path prediction of vehicle and pedestrian with the aim of collision risk prediction. The method jointly models the paths of pedestrian and vehicle in a single, unifying Dynamical Bayesian Network (DBN), consisting of two sub-graphs modeling the vehicle and the pedestrian. Contextual cues from the pedestrian, driver and environment are used to extend the prediction horizon. The subgraphs capture the awareness of pedestrian and driver to anticipate a change of motion, and consider where such stopping is likely to occur. To model potential motion coupling, the subgraphs share a latent state which observes their potential collision course. To evaluate the incremental benefits of the proposed model components, we compare six models with varying access to the proposed context cues, and collected 93 sequences covering 9 combinations of dynamics and awareness for the driver (driving/braking; (un)aware of pedestrian) and pedestrian (crossing/stopping; (un)aware of vehicle) using on-board sensors of an instrumented vehicle. Successively, collision risk is obtained by a probabilistic intersection operation on the predicted paths.

The experiments show that with prediction horizons of 1.5 s, context-aware models outperform context-agnostic models in path prediction for scenarios where either traffic participant stops, while being on par with the baselines for scenarios with continuous movement. Additionally, we show that driver-attention-aware models improve collision risk prediction capability compared to driver-agnostic models. This illustrates that driver contextual cues can support a more anticipatory collision warning and vehicle control strategy.

5

5.1. INTRODUCTION

While collision prevention and active pedestrian safety systems have provided major safety improvements over the past few decades, global fatality rates per unit of population have stayed constant between 2013 and 2017. According to the World Health Organization, pedestrians and cyclists present 26% of global traffic deaths (WHO, 2018). According to the European Road Safety Observatory, more than half of serious crashes that involve pedestrians and cyclists occur while crossing the road, and 32% occur on crossing facilities (European Commission, 2018; SWOV, 2010).

To make significant improvements in driver assistance systems like collision warning and automated emergency braking, a better understanding is needed of how the driver and nearby road users will behave in the near future. Ideally, a prediction horizon of 2.5 s is achieved, at which point the driver “feels no danger” (Winner, 2016).

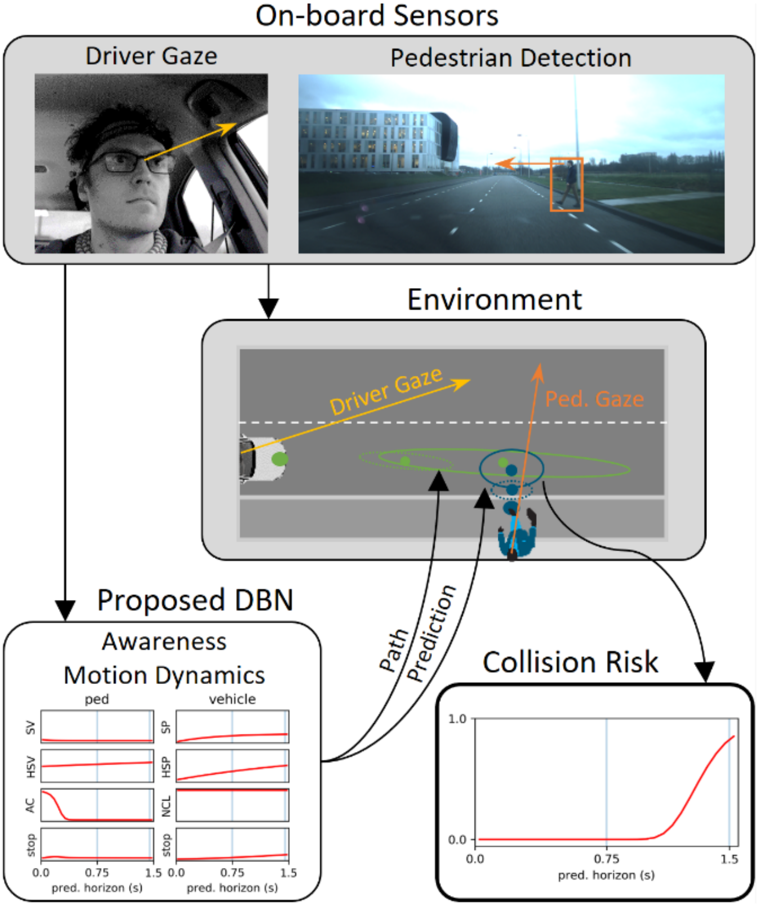


Figure 5.1: The system assesses joint awareness of pedestrian and driver in a scenario of a potentially crossing pedestrian. Cues about the driver, pedestrian and spatial environment are collected from on-board sensors. A probabilistic framework based on a Dynamic Bayesian Network (DBN) estimates latent states of awareness of the driver and pedestrian to predict their future motion. Consecutively, based on the predicted positions, future collision risk is estimated.

Changes in behavior make it hard to anticipate the future path of pedestrians, e.g., they can stop walking or change direction in an instant. This makes it particularly challenging for the ego-vehicle to deal with pedestrians which potentially cross the road. Current on-market driver assistance systems generally extrapolate the observed motion and thereby disregard such changes in their predictions. Camera-based driver monitoring systems can detect fatigue, drowsiness, distraction, gestures, signs of being drunk and readiness to take over from automated driving. On-market systems for collision warning have been employed as early as 2007 (Toyota/Lexus) by monitoring head pose and eye opening. Recent releases allow for SAE level 2 driving on specially mapped highways (Cadillac Super Cruise, 2018), in traffic jams with restricted velocity (BMW Extended Traffic Jam Assistant, 2018), or in single-lane cruising (Nissan ProPilot, 2019). Mercedes-Benz announced that the S-Class released in September 2020 employs a driver camera that will monitor driver's readiness to take over from automated driving mode on highways in an SAE level 3 system as of autumn 2021. This will enable the driver to perform non-driving related tasks for up to 10 s. Current collision risk prevention systems detect driver control actions but do not observe driver state and situation awareness to predict driver behaviour.

To better predict a possible collision, advanced driver assistance systems can evaluate clues about whether either the driver or the pedestrian is aware of the developing hazard and act accordingly. Particularly, a pedestrian with the intention to cross is less likely to do so dangerously if the pedestrian is aware of the approaching vehicle. Similarly, when the driver is aware of the crossing pedestrian, he/she is less likely to benefit from a collision warning. Driver and pedestrian awareness can be evaluated jointly (individual-aware) by simultaneously considering their awareness of the other's presence; or mutually (group-aware), e.g., assessing the driver's belief about the pedestrian's awareness of the vehicle. In this work, we focus on joint awareness. The estimation of driver and pedestrian awareness requires a basic insight in how they attend and perceive. Unfortunately, awareness is a latent state which cannot be measured directly, but gaze or head pose may provide clues on the awareness likelihood.

In this paper, we consider the setting of a potentially crossing pedestrian and an approaching vehicle which has the right of way. We present a method which uses context cues about the spatial environment, driver-pedestrian joint awareness and potential motion coupling to estimate the future paths of both participants. Successively, collision risk is assessed in a probabilistic manner. See Fig. 5.1 for an illustration of the overall system. Specifically, we extend the DBN method from Kooij *et al.* (2019), which models an individual pedestrian in a probabilistic framework based on a Dynamic Bayesian Network (DBN), and we model the interactive paths of the pedestrian and vehicle based on joint awareness of pedestrian and driver as well as environmental cues, all obtained by on-board sensors. Driver gaze and pedestrian head pose are used as awareness cues of the other road-user to anticipate their future motion dynamics. Based on the predicted situation, we estimate the collision risk which can successively be incorporated in a robust collision warning/control strategy and can be deployed in future series vehicles for improved SAE level 0 performance.

In the following section, we provide a brief overview of the components required for path prediction, including the estimation of road-user awareness, and collision risk prediction. Section 5.3 describes the context-based prediction model. Sections 5.4 and 5.5 provide a description of the obtained dataset and procedures for parameter estimation of the models, whose results are compared in Section 5.6. These results are then used to discuss the incremental benefits from context-agnostic to the driver-aware models in Section 5.7.

5.2. RELATED WORK

In this section, we discuss existing work on object localization, path prediction and collision risk prediction.

For intelligent vehicles, localization of scene objects is typically done in a 3D world coordinate frame. Noisy measurements of pedestrians and the ego-vehicle can be temporally filtered to estimate the true object state consisting of position and dynamics such as acceleration with accompanying uncertainties under motion assumptions. Pedestrians are typically localized from the ego-vehicle by on-board sensors such as stereo cameras, LIDAR and RADAR, or a combination of these modalities (Roth *et al.*, 2019; van der Sluis *et al.*, 2020). The ego-vehicle is usually localized in the 3D world coordinate frame by fusion of multiple modalities such as GPS, IMU, steering wheel angle and wheel rotation sensors.

Path prediction of traffic participants has seen great attention in the recent years. Vulnerable road user path prediction is covered in recent surveys (Ridel *et al.*, 2018; Rudenko *et al.*, 2020), while Lefèvre *et al.* (2014) cover motion prediction of vehicles. We go in-depth on context cues for path prediction and motion models.

5.2.1. CONTEXT CUES FOR PATH PREDICTION

Path prediction methods can use context cues to anticipate intentions and future behavior of traffic participants. These cues include observable gestures, poses or movements that become meaningful antecedents for behavioral changes when observed in a particular context. Rudenko *et al.* (2020) define contextual cues as "*all relevant internal and external stimuli that influence motion behavior*". Context cues can be subdivided into static and dynamic environmental cues, and object cues.

Static environmental context cues model the effect of the static environment on the agent, e.g., road topology (Pool *et al.*, 2020, 2017), or expected behaviors such as following traffic rules and preferred paths, minimizing hindrance to self and others (Wang *et al.*, 2019), and avoiding hazards.

Dynamic environmental context cues include changes in dynamics due to awareness of other traffic participants and allow for modeling adaptations in behavior. E.g., Kooij *et al.* (2014) model whether the ego-vehicle and a crossing pedestrian are on a collision course, and Pellegrini *et al.* (2009) use the expected point of closest approach to model potential collisions of multiple agents.

Object context cues incorporate features linked to the object of interest, e.g., pedestrian states like walking or standing and transitions like starting or stopping observed from per-frame postural features (Quintero *et al.*, 2015) or temporal tracking thereof. Keller and Gavrila (2014) improve pedestrian path prediction by using dense optical flow features. How pedestrians approach a crosswalk may indicate if they intend to use it or walk by (Völz *et al.*, 2019). Other examples include time-to-collision, minimum future distance, relative velocity and arm gestures, e.g., Kooij *et al.* (2019); Neogi *et al.* (2020); Pellegrini *et al.* (2009); Pool *et al.* (2020, 2017); Roth *et al.* (2016).

PEDESTRIAN-RELATED OBJECT CUES

Psychological studies show that various attributes of pedestrians may indicate awareness of oncoming traffic (Chen *et al.*, 2019): body pose, explicit and implicit gestures, facial expressions (such as smiling), assertive behavior (such as walking fast towards the crosswalk), shorter lateral distance (such as stepping on the curb instead of stopping on the sidewalk), not paying attention to traffic, or being engaged in cellular activities (Rangesh and Trivedi, 2018). Gaze is an important non-verbal cue of pedestrians being aware of on an oncoming vehicle while intending to cross, and may indicate yielding behavior depending on the vehicle's approach (speed, collision course) (Kooij *et al.*, 2019; Rasouli *et al.*, 2018). The duration of gaze also correlates with the uncertainty of the opportunity to cross (Brouwer *et al.*, 2016). However, pedestrian gaze is challenging to observe from a vehicle's perspective. Head pose provides a proxy for gaze which is easier to observe. In the intelligent vehicle domain, continuous head pose has been estimated using regression models or by head orientation specific classifiers (Flohr *et al.*, 2015; Kooij *et al.*, 2019; Ridel *et al.*, 2019).

DRIVER-RELATED OBJECT CUES

While perceptual errors contribute to 44% of driver-caused accidents, driving accidents only correlate weakly with perceptual ability (Hills, 1980). This means that while perceptual ability is an awareness prerequisite, situation awareness depends primarily on how we distribute our attention and how we interpret what we attend. Gaze provides a major cue for inferring awareness as a latent state (Wu *et al.*, 2005) and allows for driver behavior prediction (Martin *et al.*, 2018). Other cues indicative for attention include frequency of gaze fixations, dwell time, pupil size, saccades, smooth pursuit, rate of blinking, fMRI "brain mapping" and head pose (Siddharth and Trivedi, 2020; Stapel *et al.*, 2020; Wu *et al.*, 2005). Additionally, foot, hand and upper body movements form cues for driver behavior (Deo and Trivedi, 2020). While attention estimation can predict driver intention and awareness, behavioral responses provide confirmatory clues: Fukagawa and Yamada (2013) and Phan *et al.* (2014) propose methods for estimating driver awareness of pedestrians from behavioral responses measured from vehicle signals, such as accelerator pedal position, braking force and steering wheel angle. In our previous work (Roth *et al.*, 2016), we employ driver head pose to model the driver's awareness of a crossing pedestrian.

5.2.2. MOTION MODELS

Motion models differ in the way they represent, parametrize, learn and solve the task of path prediction. They can be subdivided into physics-based, pattern-based and planning-based methods (Rudenko *et al.*, 2020). Physics-based methods represent motion by explicitly defined dynamic equations of one or more underlying dynamical models. Pattern-based methods approximate arbitrary motion dynamics from training data (Alahi *et al.*, 2016; Gupta *et al.*, 2018; Keller and Gavrila, 2014; Li *et al.*, 2020; Pool *et al.*, 2020; Ridel *et al.*, 2019). Planning-based methods model long-term motion by finding path hypotheses towards a goal, e.g. the work of Lee *et al.* (2017).

In this work, we focus on physics-based models which represent the dynamic state of the road user probabilistically and allow for interpretation. Simple motion dynamics can be modeled by Linear Dynamical Systems (LDS), which commonly assume a linear relationship between states and measurements with Gaussian noise. Under these assumptions, the Kalman Filter (KF) (Welch and Bishop, 2006) is an optimal filtering algorithm, which has been widely applied for pedestrian and vehicle tracking (Lefèvre *et al.*, 2014; Schneider and Gavrila, 2013).

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 Switching LDS (SLDS). SLDS can be extended by dynamical models to incorporate contextual cues for path prediction (Kooij *et al.*, 2019; Quintero *et al.*, 2015). Li *et al.* (2020) combine the path prediction output of Kooij *et al.* (2019) with a sequence-to-sequence trajectory generation method by applying an adaptive weighting algorithm.

While these methods incorporate contextual cues to the path prediction, they only partially use cues which capture the nature of pedestrian-vehicle encounters. We now review methods which specifically model multiple traffic participants. Social force models have been introduced to model the influence of nearby vulnerable road users on each other (Alahi *et al.*, 2016). Similarly, Wang *et al.* (2019) resolve courtesy behavior between two vehicles approaching an intersection by jointly inferring the empathetic intent as respectively observed or demonstrated approaching behavior (cautious/aggressive). Gupta *et al.* (2018) model the interaction of multiple people to predict their collision free paths, based on past trajectories. Lee *et al.* (2017) propose a method to predict the paths of multiple traffic participants including static and dynamic environmental context based on RNNs. Gupta *et al.* (2019) model negotiation between pedestrians and vehicle and show in simulation results, that traffic flow improved compared to best-practice behavior of autonomous vehicles (always stop). Neogi *et al.* (2020) model pedestrian-vehicle interaction by incorporating vehicle velocity and distance to the pedestrian. Braeuchle *et al.* (2013) combine motion prediction of a pedestrian and a vehicle by defining the pedestrian state relative to the vehicle. A Bayesian network is used to find the corresponding motion model for the vehicle and the pedestrian to decide on an evasive manoeuvre for the vehicle.

These methods model interactive behavior, but neglect joint awareness cues of multiple traffic participants. Ridel *et al.* (2019) perform pedestrian path prediction using LSTMs based on past trajectories of ego-vehicle and pedestrian, and pedestrian head pose. Similarly, Kooij *et al.* (2019) use pedestrian head pose to model awareness of the ego-vehicle and influence stopping behavior. In our previous work (Roth *et al.*, 2016), we argue that interactive behavior benefits most if awareness of pedestrian as well as driver are jointly considered. This work extends the method of Kooij *et al.* (2019) by measuring driver head pose to infer driver awareness of the pedestrian. Driver awareness is incorporated to predict stopping of the ego-vehicle in case of an aware driver, or conversely to emit a collision warning in case of an inattentive driver.

5.2.3. Collision Risk Prediction

Collision risk prediction can be categorized into physical model-based and data-driven methods (Dahl *et al.*, 2019). 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, though these categories can partially overlap (Dahl *et al.*, 2019). Söntges *et al.* (2018) present a SBTM method by computing time-to-react from over-approximating reachable sets. De Nicolao *et al.* (2007) 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 (Braeuchle *et al.*, 2013) or discrete (Brouwer *et al.*, 2016) integration. Braeuchle *et al.* (2013) use a compound car-pedestrian geometric model to infer a joint spatial probability distribution. A collision risk is estimated by integration over predicted distributions for all time steps. The method of Brouwer *et al.* (2016) fuses predicted object occurrences from four pedestrian motion models in a probabilistic fusion grid. A collision risk is estimated by summation over all grid cells inside the collision corridor. In our previous work (Roth *et al.*, 2016), we 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, which is defined by all possible intersections between vehicle and pedestrian locations.

5.3. DRIVER AND PEDESTRIAN JOINT AWARENESS FOR COLLISION RISK PREDICTION

We aim for an accurate path prediction of both vehicle and pedestrian to estimate the collision risk in a lateral crossing pedestrian scenario. We predict if the pedestrian or driver is likely to yield in case the other continues in a setting where the driver has the right of way.

Our model builds upon the model by Kooij *et al.* (2019), who argue that the pedestrian's decision to continue walking or to stop is largely 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. They use this knowledge to model the pedestrian motion by a context-based SLDS (DBN), which switches between the dynamics of walking and standing depending on the inferred pedestrian's intent.

In our previous work (Roth *et al.*, 2016), we argue analogously to incorporate vehicle dynamics, and that the driver's intent will be largely influenced by whether the driver has seen the pedestrian, and whether the vehicle is on collision course with the pedestrian. We extend our previous work by further vehicle context, i.e., the distance of the vehicle to the pedestrian's crossing location, analyze different driver awareness modalities and provide a thorough analysis on path prediction and collision risk prediction.

The contributions of this work are as follows: We present a probabilistic framework for joint path prediction of vehicle and pedestrian using observed kinematics, head pose of pedestrian and driver gaze, as well as context cues of the environment. We provide a comprehensive analysis of the effect of pedestrian, vehicle and driver context-cues on the situations where an intervention of either road user is needed to avoid a collision. More specifically, we evaluate different driver monitoring modalities (invasively measured head pose Roth *et al.* (2016), estimated head pose, estimated eye gaze). We employ our framework on looking-in and looking-out onboard vehicle sensor data to predict forward collision risk.

5.3.1. DBN

We model the behavior of the pedestrian and the vehicle with two SLDSes which are jointly controlled within a single, unifying DBN. In this way, we can incorporate factors which influence each participant independently (e.g., awareness of the other), but also capture interdependencies such as being on collision-course, joint awareness or other forms of implicit communication. In this study, we limit the interdependencies to a shared collision-course latent state.

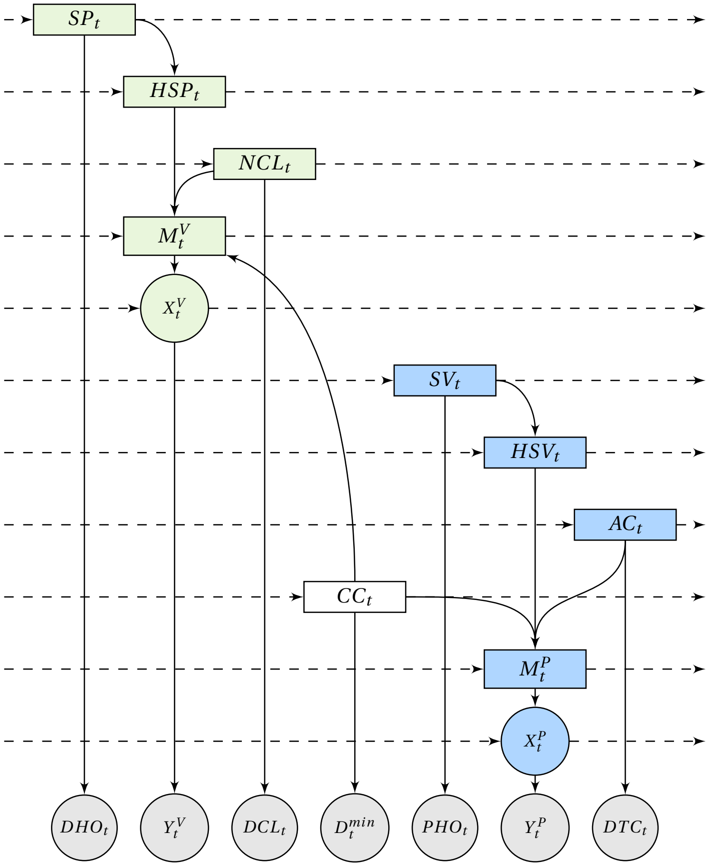


Figure 5.2: Directed graphical model representation of the Dynamic Bayesian Network (DBN). Discrete nodes are rectangular, continuous nodes are circular. Grey nodes represent observable context variables while the other nodes represent latent context states. Dashed lines depict temporal connections between 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 5.1.

The DBN structure is shown in Fig. 5.2 while node definitions and their purpose are provided in Table 5.1. We recapitulate the underlying concepts. For a thorough mathematical foundation of LDS, SLDS and DBN, we refer to Kooij *et al.* (2019).

The DBN consists of two subgraphs, one for the pedestrian and one for the vehicle, which will be described in more detail in Sections 5.3.1 and 5.3.1. The pedestrian subgraph is congruent with the DBN of Kooij *et al.* (2019). The vehicle subgraph displays analogous behavior for the vehicle, by encoding driver awareness by driver gaze and braking manifestation by being close to the crossing line of the pedestrian.

PEDESTRIAN-RELATED CONTEXT STATES

The pedestrian P can exhibit one of two motion types: *walking* ($M_t^P = m_w$, constant velocity) and *standing* ($M_t^P = m_s$, constant position). The motion state of the pedestrian contains

Table 5.1: Latent context states, their associated observation and the purpose within the DBN structure States are grouped by driver, vehicle, pedestrian and shared contexts.

Latent State	Abbr.	Observable	Abbr.	Purpose
driver-sees-pedestrian	SP	driver-head-orientation (gaze)	DHO	encodes driver's awareness of the pedestrian
has-seen-pedestrian	HSP	-	-	memorizes driver's (past) awareness of the pedestrian
vehicle-near-crossing-line	NCL	distance-to-crossing-line	DCL	manifests typical location of braking
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	SV	pedestrian-head-orientation	PHO	encodes pedestrian's awareness of the driver/vehicle
pedestrian-has-seen-vehicle	HSV	-	-	memorizes pedestrian's (past) awareness of the driver
pedestrian-at-curb	AC	pedestrian-distance-to-curb	DTC	manifests typical location of stopping
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

5

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:

$$A^{(m_w)} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, A^{(m_s)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.1}$$

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

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 $M_{t+1}^P = \text{Cat}\left(M_t^P, AC_{t+1}, HSV_{t+1}, CC_{t+1}\right)$ as shown in Figure 5.2.

The pedestrian awareness context SV_t models whether the pedestrian sees the approaching vehicle. Head orientation PHO_t forms the evidence.

The context variable HSV_t memorizes whether the pedestrian has seen the vehicle in the past, acting as a logical *OR* between previous HSV_{t-1} and current SV_t.

The environment context AC_t 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 $X_t^V = \left[x_t, y_t, \dot{x}_t, \dot{y}_t\right]^T$. It uses a constant velocity model while driving, and a velocity decay model for braking:

$$A^{(m_d)} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, A^{(m_b)} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & d & 0 \\ 0 & 0 & 0 & d \end{bmatrix} \tag{5.2}$$

The decay parameter $d < 1$ is selected to represent a velocity half-life of 0.5 s. Also, the vehicle V observes its own velocity, resulting in the observation matrix $C^V = \left[\begin{smallmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{smallmatrix}\right]$.

For the vehicle, the context-based SLDS' switching state M^V is encoded as a categorical distribution $M^V_{t+1} = \text{Cat}\left(M^V_t, NCL_{t+1}, HSP_{t+1}, CC_{t+1}\right)$.

The driver awareness context SP models the driver's awareness of the pedestrian. It is inferred from the attention eccentricity, i.e., the absolute visual angle difference between the driver's center of gaze (or head direction) and the pedestrian, DHO_t.

The context variable HSP_t memorizes whether the driver has seen the pedestrian analogous to HSV_t.

The static environment context NCL_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 trajectories with their momentary estimated velocities (Pellegrini *et al.*, 2009).

5

5.4. PARAMETER ESTIMATION

The model contains parameters for motion dynamics, context variables, such as priors and transition probabilities, as well as context observation distributions.

We find model parameters by data-driven initialization, followed by gradient based optimization. Both use the dataset we introduce in Section 5.5.

5.4.1. MODEL PARAMETER INITIALIZATION

LDS PARAMETERS

The underlying motion models of M^V and M^P are represented by LDSes which model process noise Q and observation uncertainty R.

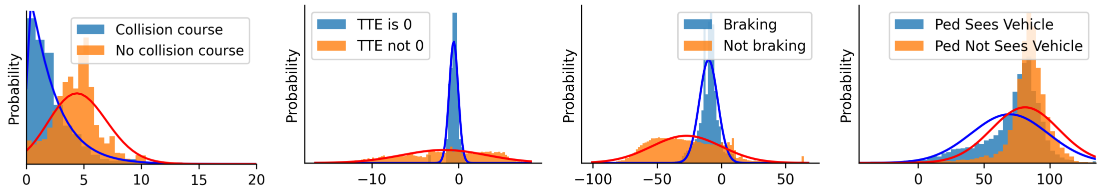
Process noise Q of vehicle and pedestrian are set for both position and velocity states and are limited to the main diagonal. Values were selected to reflect model uncertainty under typical velocity changes of drivers and pedestrians (Nazir *et al.*, 2014; Saptoadi, 2017).

Observation noise R is set to reflect typical variance of measurement noise for pedestrian detection and vehicle movement observed in our research platform, see Section 5.5.

CONTEXT STATES AND TRANSITION PROBABILITIES

The relations between binary context states Z are described by transition probabilities while the influence of continuous context observations e_t depends on the likelihoods derived from the distributions $P(e_t|Z_t = \text{True})$ and $P(e_t|Z_t = \text{False})$. These distributions and transition matrices need to be estimated. As in Kooij *et al.* (2019), the distributions are not learned through training, but found through fitting on the database of Section 5.5. The annotation value of each context state ($Z_t = \text{True}$ or $Z_t = \text{False}$) is defined by scenario (PHO, DHO), manual event annotation (DTC, DCL) or manually setting a parameter threshold guided by distributions (D^{min}).

The resulting distribution parameters are obtained by Maximum-Likelihood-Estimation of Gaussian (DTC, DCL), Gamma (D^{min}, DHO) or von-Mises distribution (PHO). The context observation distributions fitted on our dataset (Section 5.5) are illustrated in Fig. 5.3.

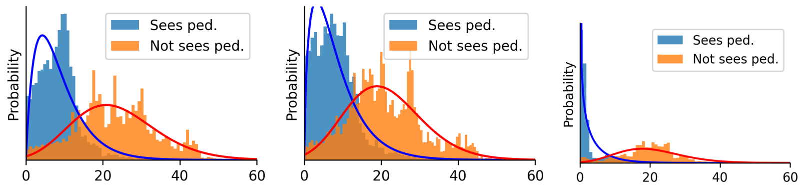


(a) D^{min}: Minimum distance along approach (m)

(b) DTC: Pedestrian distance to curb (m)

(c) DCL: Vehicle distance to crossing line (m)

(d) PHO: Pedestrian head orientation (deg)



(e) DHO: Attention eccentricity (measured head pose) (deg)

(f) DHO: Attention eccentricity (estimated head pose) (deg)

(g) DHO: Attention eccentricity (estimated eye gaze) (deg)

Figure 5.3: Histograms and the fitted distributions of the context observations. The histograms represent the observation in the dataset for the respective ground truth class. The distribution functions are fitted on the underlying data. See Section 5.4.1.

For the collision-course context variable CC, we chose a minimum future distance value of $D^{min} < 2.6\,\mathrm{m}$ for being on a collision course, based on the histograms of Figure 5.3a.

Context transition probabilities are found through enumeration. There are two exceptions on this procedure: (a) estimated absolute driver attention eccentricity (Fig. 5.3g), where the histograms suggest a stricter discrimination criterion of 4° compared to 8° as suggested by the fitted distribution. (b) pedestrian head orientation (Fig. 5.3d), where attentive sequences contained many observations where head pose was incorrectly classified in the direction of walking. Attentive poses were thus limited to head angles directed towards the vehicle in a range from -30° to 30°.

5.4.2. MODEL PARAMETER OPTIMIZATION

To optimize model parameters, we employed the parameter estimation method of Pool *et al.* (2020), which makes use of error back propagation. We maximize the observation log likelihood of the vehicle and pedestrian under their respective predicted Gaussian distributions, see Eq. (5.4). All intermediate time-steps up to the prediction horizon are incorporated into the loss function to enforce a consistent path. Measurements with time-to-event (TTE) $\in [-2.5\,\mathrm{s}, 3.0\,\mathrm{s}]$ are considered for optimization, to cover periods of typical motion dynamics. Missing intermediate measurements are ignored for optimization. TTE is defined in section

5.5.

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. We also enforce the latter 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: LDS process noises (Q) of pedestrian and vehicle, transition probabilities, and context observation distribution parameters.

5.5. DATASET

The dataset comprises of an approaching vehicle which has the right of way and a potentially crossing pedestrian. It contains scenarios with the following expected 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/driving.
- Unaware participants continue walking/driving.

Further, it contains an anomalous scenario where the pedestrian does not comply with the above attributes, by crossing while being aware of the approaching vehicle on collision course.

5.5.1. SCENARIOS

For the purpose of model parameter estimation and evaluation, vehicle-pedestrian encounters 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 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 5.4.

Scenarios 1 to 4 represent non-collision-course scenarios, meaning the pedestrian has sufficient time to cross. 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 optimization. The remaining scenarios (5-8) are safe through a change in behavior by either the driver or pedestrian due to awareness of the other participant.

A total of 93 sequences were recorded with 4 trained drivers and 4 pedestrians. Each scenario was captured by 8 to 20 sequences, see Table 5.2. The time between the first pedestrian detection and the pedestrian reaching the curb is (min / max / mean = 1.3 s / 3.2 s / 2.9 s). In that period, the pedestrian detection recall was 83 %.

Table 5.2: Number of sequences per scenario.

Scenario	1	2	3	4	5	6	7	8	9[a]
# sequences	8	13	8	8	10	8	20	10	8

5.5.2. INSTRUMENTATION

All data was collected with a TU Delft experimental vehicle, whose instrumentation is described in further detail in Ferranti *et al.* (2019). Vehicle position, orientation and velocity are obtained from an odometry system which fuses differential GPS with GNSS, IMU, visual odometry from stereo vision, steering wheel angle and wheel ticks. 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, 10 fps, 1936 × 1216 px) mounted near 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 was recorded with two systems. *Estimated* eye gaze and head pose were recorded with a high-end commercial off-the-shelf eye tracker (Smarteye: 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 (Roth *et al.*, 2016; Roth and Gavrila, 2019). Additionally, the driver was observed by a camera mounted above the speedometer for visual verification purposes. All sensor data was spatially calibrated and resampled to a target rate of 20 Hz.

3D pedestrian positions were obtained by 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.* (2019). (2) Depth was found by median stereo disparity (Hirschmüller, 2008) of the 2D bounding box. (3) 3D positions were gained by ego motion compensation of the vehicle. To obtain an indication of depth sensing noise, we manually annotated the pedestrian's path and calculated the standard deviation of the difference between the depth estimate and the detection's projection onto the annotated path. This resulted in a standard deviation of 1.79 m on the used dataset (Section 5.5).

Similarly to Kooij *et al.* (2019) and Roth *et al.* (2016), we infer pedestrian's focus-of-attention from pedestrian head orientation. We apply the method of Braun *et al.* (2016) 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. 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 TTE, for each time-step of each sequence (negative TTE: before event). We employ map information and ego-vehicle localization to estimate the location of the curb side.

5.5.3. PROCEDURES

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, 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.

5.6. RESULTS

5

To evaluate the incremental benefits of the DBN model components for an intelligent collision warning system, we compare six models with varying access to the used context cues on their joint prediction performance of vehicle and pedestrian behavior and collision risk.

We adopt two evaluation metrics: the ability to predict driver and pedestrian location 1.5 s into the future, and collision risk across multiple prediction horizons. To accommodate the small size of the database, evaluation is performed using 5-fold cross validation.

5.6.1. EVALUATION METRICS

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

PATH PREDICTION PERFORMANCE

Two performance metrics are used to evaluate path prediction performance: (a) Euclidean distance error between predicted expected position and future position estimation $\widetilde{X}_{t+t_p}$:

$$\text{error}\left(t_p|t\right) = \left|\mathbb{E}\left[\widetilde{P}_{t\to t+t_p}(X_t)\right] - \widetilde{X}_{t+t_p}\right| \tag{5.3}$$

and (b) the log likelihood of the future position estimation $\widetilde{X}_{t+t_p}$ under the predictive distribution:

$$\text{loglik}\left(t_p|t\right) = \log\left[\widetilde{P}_{t\to t+t_p}\left(\widetilde{X}_{t+t_p}\right)\right] \tag{5.4}$$

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

COLLISION RISK

We determine the probability for a collision 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 $\widetilde{P}_{t\to t+t_p}(X_t) = \mathcal{N}(\mu_{t\to t+t_p}, \sigma^2_{t\to t+t_p})$ be the predictive position of either pedestrian P or vehicle V. The *combined* predictive position is then defined as $\widetilde{P}^{\phi}_{t\to t+t_p}(X^{\phi}_t) = \mathcal{N}(\mu^P_{t\to t+t_p} - \mu^V_{t\to t+t_p}, (\sigma^P_{t\to t+t_p})^2 + (\sigma^V_{t\to t+t_p})^2)$. The collision risk predicted from t for $t+t_p$ is given by:

$$\mathrm{CR}\left(t_p | t\right) = \int_{A^{\phi}} \widetilde{P}^{\phi}_{t\to t+t_p}(X^{\phi}_t) \tag{5.5}$$

with A^{ϕ} being the combined spatial extent of vehicle and pedestrian dimensions combined. This resembles the collision risk estimation method of Braeuchle *et al.* (2013).

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. We define a collision as any instance where (1) the scenario represents a collision scenario and (2) the *LDS* estimates the momentary (i.e., $t_p = 0$) collision risk to be 5% or larger. In order to assess the collision risk prediction performance at various prediction horizons, we select a fixed false positive rate (FPR) and find the attainable true positive rate (TPR) for each prediction horizon t_p.

5

5.6.2. MODEL VARIANTS

We evaluate four context-aware models including the method of Kooij *et al.* (2019) which differ in their access to pedestrian, vehicle and driver context, and compare them to two context-agnostic models. An overview of the used context cues of the models is given in Table 5.3. All models were optimized individually as described in Section 5.4.

CONTEXT-AGNOSTIC MODELS

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

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/VEHICLE/DRIVER CONTEXT

We analyze four variants of the model presented in Figure 5.2 which take different amounts of context into account: *DBN.p* represents the context-based pedestrian path prediction method of Kooij *et al.* (2019). The method is driver-agnostic and models the vehicle dynamics as a context-agnostic SLDS. *DBN.pv* is vehicle-aware and extends *DBN.p* with vehicle static environment cues but remains driver-agnostic. It includes proximity of the vehicle to the crossing line of the pedestrian (NCL). *DBN.pvh* additionally uses driver head pose as an awareness cue (SP). *DBN.pvg* uses driver eye gaze instead of driver head pose.

5.6.3. PATH PREDICTION

Table 5.4 depicts *loglik* and Euclidean distance error of both pedestrian and vehicle for prediction horizon $t_p = 1.5\,\mathrm{s}$ averaged over periods where typical changes in dynamics occur (pedestrian: TTE $\in [-0.5\,\mathrm{s}, 2.0\,\mathrm{s}]$, vehicle: TTE $\in [-0.5\,\mathrm{s}, 3.0\,\mathrm{s}]$; TTE ranges define times where predictions are made for).

Table 5.3: Context cues and number of motion models per road user used in the models. DBN suffixes denote used context: p: pedestrian (Kooij *et al.*, 2019); v: vehicle (NCL); h: driver head pose; g: driver gaze.

Context cue	LDS	SLDS	DBN.p	DBN.pv	DBN.pvh	DBN.pvg
Ped. at-curb	-	-	x	x	x	x
Ped. awareness	-	-	x	x	x	x
Collision course	-	-	x	x	x	x
Veh. near-cross	-	-	-	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

Scenarios 1 to 4 represent non-collision course scenarios where there is no risk of collision when driver and pedestrian continue their paths. All models show similar *loglik* and Euclidean distance error for both the pedestrian and the vehicle. This is to be expected, as they denote no change in dynamics of either road user. One exception is the *SLDS* model which achieved generally lower pedestrian *loglik* performance on all scenarios. Inspection of individual runs revealed that the pedestrian observation noise causes multiple switches between walking and standing motion model, leading to erroneous predictions.

In scenarios 5 and 6, the pedestrian stops to prevent a collision. Both context agnostic models (*LDS*, *SLDS*) show poor pedestrian path prediction performance. Introducing pedestrian awareness context (all DBN models) raises the pedestrian *loglik* by approximately 2.9 compared to the *LDS* on scenario 6, which corresponds to an increased likelihood by a factor of $e^{2.9} > 18$ and reduces positional error by approximately 13 cm. The effect of adding vehicle context (*DBN.pv*) and driver context (*DBN.pvh*, *DBN.pvg*) for the pedestrian is negligible, as those contexts do not affect motion dynamics of the pedestrian. Vehicle path prediction drops in performance for scenario 5 when vehicle context and driver context is added. This is because the driver-aware DBN variants also consider the option that an aware driver might yield despite having right of way.

In scenario 7, the aware driver slows down for an unaware crossing pedestrian. The *LDS* model wrongly predicts the vehicle to continue, resulting in low vehicle *loglik*. The *SLDS* model adapts more quickly to the change of vehicle motion dynamics leading to a *loglik* of 3.0. The DBN variants further increase the *loglik* of the vehicle, and manifest in a 32 cm smaller Euclidean error for *DBN.pv* compared to the *SLDS* and a larger *loglik* when including driver awareness (*DBN.pvh* and *DBN.pvg* vs. *DBN.pv*). There is no practical difference in path prediction performance between *DBN.pvh* and *DBN.pvg*. Figure 5.4 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 ($TTE < -0.5$ s), the three models show similar performance. As the vehicle slows down, both *LDS* and *SLDS* increase in spread over various runs (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 stopping) compared to the *LDS*. The *DBN.pvg* model anticipates the change in motion dynamics resulting in a higher *loglik* and less uncertainty than the context-agnostic models.

Table 5.4: 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 $t_p = 1.5\,\mathrm{s}$. The top half of the table captures prediction performance for the pedestrian while the bottom half displays vehicle path prediction performance. See Section 5.6.2 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 traffic participant.

Scen.	CC	Ped. stops	Ped. SV	Veh. stops	Driver SP	LDS	SLDS	DBN p	DBN pv	DBN pvh	DBN pvg	LDS	SLDS	DBN p	DBN pv	DBN pvh	DBN pvg
						Pedestrian 1.5 s *loglik*						Pedestrian 1.5 s Euclidean error (cm)					
1	0	0	0	0	0	**−3.7**	−4.3	−4.1	−4.1	−4.1	−4.2	251	**214**	243	243	242	243
2	0	0	0	0	1	**−3.9**	−7.9	−4.8	−4.8	−4.8	−4.8	289	**256**	292	292	293	292
3	0	0	1	0	0	**−4.1**	−6.3	−4.6	−4.6	−4.6	−4.6	256	**246**	279	279	279	279
4	0	0	1	0	1	**−3.3**	−3.3	−3.9	−3.9	−3.8	−3.9	**170**	177	192	191	182	192
5	1	1	1	0	1	−7.3	−7.9	**−3.6**	−3.7	−3.7	−3.7	202	180	**170**	173	173	174
6	1	1	1	0	0	−6.3	−6.1	−3.4	−3.4	**−3.4**	−3.4	164	168	151	152	**151**	151
7	1	0	0	1	1	**−3.1**	−9.3	−3.7	−3.7	−3.7	−3.7	**140**	146	155	155	155	156
8	1	0	0	0	0	**−3.1**	−5.1	−3.7	−3.7	−3.7	−3.7	158	**156**	173	173	173	174
9[a]	1	0	1	0	0	**−3.2**	−10.1	−3.9	−3.9	−3.9	−3.9	201	**201**	215	215	215	215
all non-anomalous *CC* = 1 (5 to 8)						−4.0	−7.1	−3.6	−3.6	**−3.6**	−3.7	155	**155**	158	159	159	160
						Vehicle 1.5 s *loglik*						Vehicle 1.5 s Euclidean error (cm)					
1	0	0	0	0	0	−1.1	−1.2	−1.4	−1.1	**−1.1**	−1.1	55	55	**47**	55	53	51
2	0	0	0	0	1	−1.1	−1.2	−1.4	−1.1	−1.1	**−1.1**	61	63	**50**	61	61	60
3	0	0	1	0	0	−1.0	−1.2	−1.4	−1.0	**−0.9**	−1.0	49	53	**40**	50	48	46
4	0	0	1	0	1	−1.1	−1.3	−1.4	−1.2	**−1.1**	−1.1	64	67	**56**	64	63	63
5	1	1	1	0	1	**−1.0**	−1.2	−1.4	−1.7	−1.6	−1.6	**50**	55	50	122	108	117
6	1	1	1	0	0	**−0.9**	−1.1	−1.4	−1.5	−1.0	−1.1	44	52	**41**	106	52	51
7	1	0	0	1	1	−8.3	−3.0	−2.8	−2.8	−2.7	**−2.7**	248	191	196	159	**156**	160
8	1	0	0	0	0	**−1.0**	−1.1	−1.4	−1.3	−1.0	−1.0	46	47	**41**	80	46	45
9[a]	1	0	1	0	0	**−0.9**	−1.2	−1.4	−1.3	−0.9	−1.0	38	45	**36**	79	41	41
all non-anomalous *CC* = 1 (5 to 8)						−4.9	−2.2	−2.1	−2.2	−2.0	**−2.0**	154	126	125	130	**113**	115

Scenario 9[a] is anomalous as the pedestrian crosses despite seeing the vehicle. For pedestrian path prediction performance, the DBN variants show a slightly lower *loglik*, as they take into account that the pedestrian might stop due to seeing the vehicle. With respect to vehicle path prediction, scenario 9[a] shows a similar performance to scenario 8 for all models.

5.6.4. COLLISION RISK PREDICTION

We first compare 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 PREDICTION

Figure 5.5a 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. For a prediction horizon $t_p = 0.75\,\mathrm{s}$, all models predict a negligible collision risk (dashed lines). Predicting $t_p = 1.5\,\mathrm{s}$ into future, the *LDS* and *SLDS* models anticipate a collision risk of 76% and 49% respectively while the *DBN.pvg* model keeps a collision risk below 20% throughout

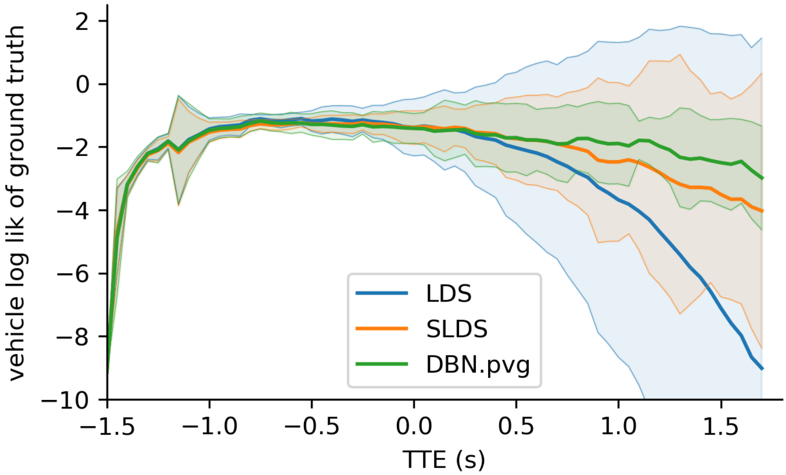


Figure 5.4: *Loglik* and standard deviation over time for a braking vehicle (scenario 7) for a prediction horizon $t_p = 1.5$ s, and drawn at the moment for which the prediction was created. The vehicle initiates braking for the crossing pedestrian between −1.8 s and 0.6 s, with most vehicles braking from 0.0 s onward.

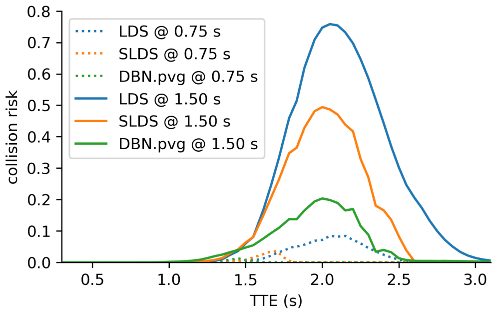
the sequence.

Figure 5.5b 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* represents all instances labelled as a collision in accordance with Section 5.6.1. All compared models (*LDS*, *SLDS*, *DBN.pvg*) depict similar maxima of collision risk within the collision window for both prediction horizons. With increasing prediction horizon, each model becomes less certain, resulting in a lower predicted collision risk value. The maxima are above 30% within the collision window for the exemplarily depicted sequence. Figure 5.5 further shows that only for *DBN.pvg*, there exists a range of collision risk thresholds (20%–33%) for which a collision warning is triggered in the collision sequence (Figure 5.5a) but not in the non-collision sequence (Figure 5.5a).

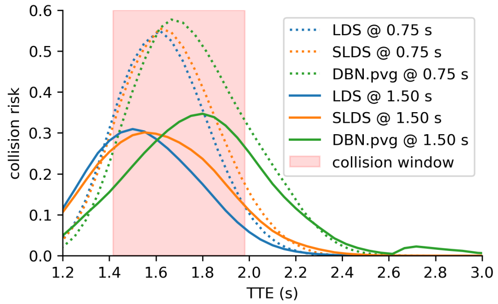
OVERALL COLLISION RISK PREDICTION

To examine how collision risk prediction performance changes with prediction horizon t_p, we select a FPR of 1% and evaluate the attainable TPR as a function of t_p. Figure 5.6 shows that for a prediction horizon up to 0.8 s, all models except the LDS achieve a TPR close to 1.0. At prediction horizons > 1 s, the TPR of the *LDS* drops below 20%.

Adding context successively results in better collision prediction for larger prediction horizons. The *DBN.pvh* and *DBN.pvg* models maintain the highest TPR for larger prediction horizons. The vehicle-agnostic *DBN.p* and driver-agnostic *DBN.pv* models perform in between the driver-aware (*DBN.pvg*, *DBN.pvh*) and the context-free (*LDS*, *SLDS*) models. At a prediction horizon of 1.5 s, The pedestrian-aware model *DBN.p* of Kooij *et al.* (2019) maintains a 30% TPR while *DBN.pvg* achieves 59%.



(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 5.5: Collision risk predictions obtained from different models for a braking vehicle (top) and collision (bottom) 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.

5

5.6.5. DRIVER MONITORING MODALITY

While driver gaze monitoring may provide a sensitive context cue for attention, accurate eye tracking can be expensive. More affordable systems such as head pose trackers may provide sufficient discrimination for collision warning support. Figure 5.7 shows a comparison between driver gaze (*DBN.pvg*) and driver head pose (*DBN.pvh*) as contextual cue for SP (sees-pedestrian). For $SP = 1$, driver gaze provides higher classification confidence in HSV (has-seen-vehicle) compared to head pose. For $SP = 0$, both models incorrectly believe that the driver has seen the pedestrian for a small fraction of sequences, though *DBN.pvg* has fewer of such sequences at all times, thus maintaining a better accuracy compared to *DBN.pvh*. However, this classification accuracy did not yield a better vehicle path prediction performance when comparing *DBN.pvg* to *DBN.pvh* in Table 5.4. Both models yielded better vehicle path prediction performance compared to the driver-agnostic model (*DBN.pv*).

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

5.7. DISCUSSION

The evaluated DBN variants perform joint path prediction and collision risk prediction for a pedestrian and vehicle in scenarios with a laterally crossing pedestrian. To evaluate the incremental benefits of the *DBN* model components, we compare six model variants with varying access to the used context cues.

Overall, our findings show that in scenarios with motion transitions of either the vehicle or the pedestrian, the context-aware models (*DBN.p*, *DBN.pv*, *DBN.pvh*, *DBN.pvg*) outperform the context-agnostic baseline models (*LDS*, *SLDS*) in terms of vehicle and pedestrian path prediction performance (cf. Table 5.4 pedestrian scenarios 5 and 6; vehicle scenario 7). The context-aware models perform on-par or better for non-anomalous collision course scenarios

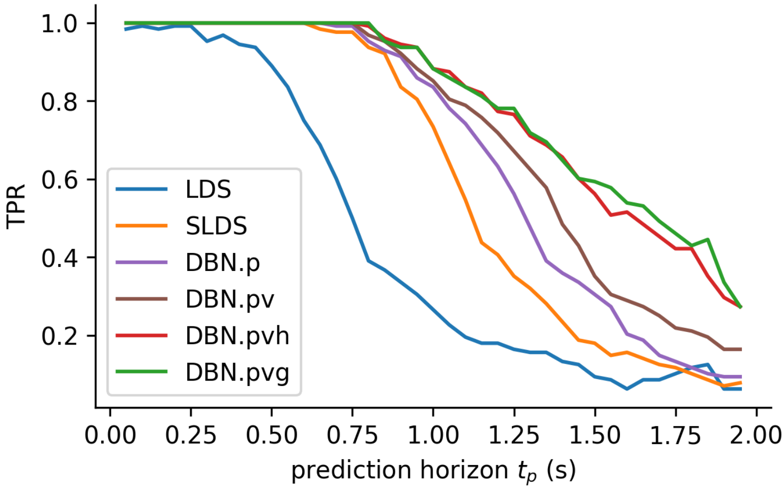


Figure 5.6: Collision risk TPR of different models obtained under a 1% FPR for various prediction horizons. Higher values denote better performance.

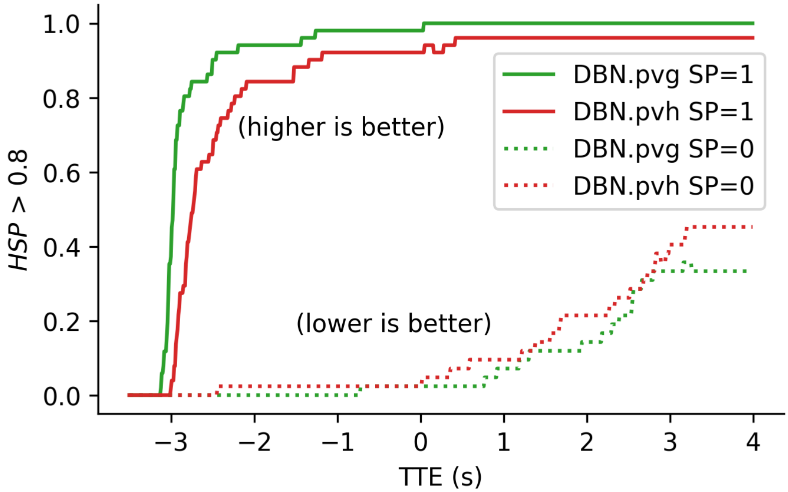
5

(cf. Table 5.4 bottom row of pedestrian and vehicle). Pedestrian context (*DBN.p*), vehicle context (*DBN.pv*) and driver awareness (*DBN.pvh*, *DBN.pvg*) provided incremental improvements in collision risk prediction performance.

When integrating pedestrian context (*DBN.p*), pedestrian path prediction performance improved (cf. Kooij *et al.* (2019)). Providing knowledge about crossing location NCL (*DBN.pv*) improved vehicle path prediction Euclidean error. The *loglik* did not improve, likely because the moment of braking onset varied by 2.4 s across sequences, which limits the accuracy gain from NCL. *DBN.pv* resulted in better collision prediction performance compared to *DBN.p*. Incorporating driver awareness (*DBN.pvg*, *DBN.pvh*) yielded a further increase of both vehicle path prediction and collision risk prediction performance (cf. Table 5.4 and Figure 5.6).

In contrast to our expectations, measuring driver gaze yielded similar path prediction and collision prediction performance compared to measuring driver head pose. This might suggest that cheaper monitoring techniques may be sufficient to achieve the prediction advantages of Figure 5.6. However, when multiple road users or driving distractions are introduced, it is likely that driver awareness will be disambiguated 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 (Stapel *et al.*, 2020), though such evaluations would require natural as opposed to instructed viewing behavior, and other spatial regions competing for attention.

The driver-aware models perform joint path prediction of two road users with interaction encoded in part via a shared latent state. This indicates that using CC (collision-course), which captures joint but not mutual awareness, cannot explicitly disambiguate situations where both driver and pedestrian are attentive. The models encode the following: if one road user



5

Figure 5.7: Classification performance of *DBN.pvg* and *DBN.pvh* on the hidden *HSP* state on sequences where driver is instructed to be attentive ($SP = 1$) and inattentive ($SP = 0$).

A is aware of the other *B*, it induces a change in dynamics of *A* which influences the shared collision course context node, which in turn participates to change in motion dynamics of *B*. The limitations of the current structure lie in modeling a direct path of influence of awareness of *A* to the change of dynamics of *B*. Considered further, mutual awareness as suggested by Wang *et al.* (2019) and negotiations over the right of way would involve further aspects of game theory, i.e., modeling the others awareness of oneself. The underlying DBN provides a versatile structure to model expert knowledge. To that end, it can be extended by additional cues, such as adding a stronger structural coupling to the make motion model switches of road user A also dependent on road user B's awareness of A.

Anomalous situations which defy the anticipated behaviors, but still occur in real-world traffic (e.g., scenario 9[a]), also provide a challenge to a context-aware system. They might contradict the expert knowledge encoded by design choices such as topology and choice of context. Fortunately, the probabilistic modeling allows for softer decisions, despite favoring the designed observations: the switch of motion dynamics not only depends on the preconditioning context, but also on the current positional observations. RNNs (Pool *et al.*, 2020) might mitigate the problem of imprecise expert knowledge by learning representations from data. There is still the problem of availability of scarce data and the bias of parameter optimization towards more prominent situations in supervised learning.

In many real-world pedestrian-vehicle encounters, no transition in behavior takes place, e.g., the pedestrian has sufficient time to cross or does not have the intention to cross. The interesting and challenging part are the few time instances with sudden changes in dynamics, which a robust model needs to anticipate in order to provide a good prediction. This poses an imbalance of data toward situations where no interaction is needed. Care should be taken to deal with this imbalance when training and evaluating. E.g., we evaluated the path prediction performance over a subset of time where changes in dynamics might take place to emphasize the effect on more challenging instance of time.

Our method models joint awareness of two traffic participants. In principle, the model can incorporate an arbitrary number of traffic participants. However, the shared latent states increase polynomially by the number of road users, which makes modeling a large number of interactive behaviors computationally intractable, especially when mutual awareness is the goal (see above). Additionally, the underlying motion models are restricted to linear dynamics which might not hold true for longer prediction horizons and do not consider preferred paths. Pattern-based motion models like social forces (Alahi *et al.*, 2016) or trajectory matching (Keller and Gavrila, 2014) could be incorporated to encounter for that shortcoming. We see a gap between methods which analyze joint awareness, like the present work, and methods that deal with groups, social interactions and preferred paths. As a solution, we believe a useful model considers social interactions, as well as context cues which reflect mutual awareness. Both social interactions and preferred paths could be integrated into the DBN framework by adding context cues and altering motion models to respect preferred paths.

We estimate collision risk from predictive state distributions of the vehicle and the pedestrian. This poses the dilemma of deciding on a collision warning strategy based on the predicted collision risk value. The concrete choice gives a trade-off between having a 'safe' system (high TPR) and a 'conservative' system (low FPR). The prediction horizon analysis of Figure 5.6 allows for selecting the system which fits the application's needs. Having collision risk analytically defined based on the predictive distributions allows for integration into other collision avoidance schemes, such as evasive maneuvers, which evaluate different future paths of the ego-vehicle to base a decision for evasive trajectory. Contrary, methods which directly estimate a collision risk value based on past motion, e.g., De Nicolao *et al.* (2007), make the integration into a collision avoidance system more complex, as they do not allow for evaluating of different future paths of the road users.

5.8. CONCLUSIONS

We presented a novel method for path prediction of vehicle and pedestrian with the aim of probabilistic collision risk prediction in scenarios with a lateral crossing pedestrian. The method modeled the behavior of pedestrian and the vehicle with two Switching Linear Dynamical Systems (SLDS) which were jointly controlled in a single, unifying Dynamical Bayesian Network (DBN) using different pedestrian, vehicle and driver context cues. Successively, collision risk was computed by a probabilistic intersection operation. Overall, this work closed the loop between on-board sensors up to collision warning.

We evaluated the incremental benefits of pedestrian, vehicle and driver context in six models with varying access to the used context cues, namely Linear Dynamical System (*LDS*, one motion model), *SLDS* (two motion models), *DBN.p* (pedestrian aware), *DBN.pv* (driver-agnostic), *DBN.phg* (driver-gaze as awareness cue) and *DBN.pvh* (driver head pose as awareness cue).

For a prediction horizon of 1.5 s, the driver-aware model *DBN.pvg* outperformed context-agnostic *LDS* and *SLDS* for scenarios with a stopping pedestrian with a pedestrian *loglik* of $-6.3/-6.1/-3.4$ (*LDS*/*SLDS*/*DBN.pvg*) while being on-par for vehicle path prediction. Similarly, for the scenario with a braking vehicle, *LDS*/*SLDS*/*DBN.pvg* yielded vehicle path prediction *logliks* of −8.3/−3.0/−2.7. For scenarios of continuing motion, the models compared on-par. The collision risk warning TPR was raised from 30% (pedestrian-aware model *DBN.p* of Kooij *et al.* (2019)) to 59% for *DBN.pvg* for a prediction horizon of 1.5 s and a false positive rate (FPR) of 1% over the dataset.

Pedestrian context (*DBN.p*) and driver awareness (*DBN.pvh*, *DBN.pvg*) provided incremental improvements in path prediction and collision risk prediction performance in scenarios where either the driver brakes or the pedestrian stops. We take this as an evidence of the superior capability of anticipating future paths from joint awareness compared to simpler individual-aware and context-free baselines.

Future work involves extensions to path prediction of multiple traffic participants and mutual awareness, e.g., assessing the driver's belief about the pedestrian's awareness of the vehicle.

ACKNOWLEDGMENT

We thank Ewoud Pool for sharing his code and expertise on DBN. We also thank Markus Braun for obtaining pedestrian detections and head pose on our database.

This project was partially funded by the NWO-TTW Foundation, the Netherlands, under the project "From Individual Automated Vehicles to Cooperative Traffic Management (IAVTRM)", STW #13712.

5

5.9. REFERENCES

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.

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).

M. Braun, S. Krebs, F. Flohr, and D. Gavrila, *EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes,* IEEE Trans. on Pattern Analysis and Machine Intelligence **41** (8), pp. 1844–1861 (2019).

M. Braun, Qing 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.

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.

W. Chen, X. Zhuang, Z. Cui, and G. Ma, *Drivers' recognition of pedestrian road-crossing intentions: Performance and process,* Transportation Research Part F: Traffic Psychology and Behaviour **64**, pp. 552–564 (2019).

J. Dahl, G. R. de Campos, C. Olsson, and J. Fredriksson, *Collision Avoidance: A Literature*

Review on Threat-Assessment Techniques, IEEE Trans. on Intelligent Vehicles **4** (1), pp. 101–113 (2019).

G. De Nicolao, A. Ferrara, and L. Giacomini, *Onboard Sensor-Based Collision Risk Assessment to Improve Pedestrians' Safety,* IEEE Trans. Veh. Technol **56** (5), pp. 2405–2413 (2007).

N. Deo and M. M. Trivedi, *Looking at the Driver/Rider in Autonomous Vehicles to Predict Take-Over Readiness,* IEEE Trans. on Intelligent Vehicles **5** (1), pp. 41–52 (2020).

European Commission, *Pedestrians and Cyclists,* European Commission, Directorate General for Transport , pp. 1–44 (2018).

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.

F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila, *A Probabilistic Framework for Joint Pedestrian Head and Body Orientation Estimation,* IEEE Trans. on Intelligent Transportation Systems **16** (4), pp. 1872–1882 (2015).

Y. Fukagawa and K. Yamada, *Estimating Driver Awareness of Pedestrians from Driving Behavior Based on a Probabilistic Model,* in *IEEE Intelligent Vehicles Symposium (IV)* (2013) pp. 1155–1160.

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.

S. Gupta, M. Vasardani, and S. Winter, *Negotiation between Vehicles and Pedestrians for the Right of Way at Intersections,* IEEE Trans. on Intelligent Transportation Systems **20** (3), pp. 888–899 (2019).

B. L. Hills, *Vision, Visibility, and Perception in Driving,* Perception **9** (2), pp. 183–216 (1980).

H. Hirschmüller, *Stereo Processing by Semi-Global Matching and Mutual Information,* IEEE Trans. on Pattern Analysis and Machine Intelligence **30** (2), pp. 328–341 (2008).

C. G. Keller and D. M. Gavrila, *Will the Pedestrian Cross? A Study on Pedestrian Path Prediction,* IEEE Trans. on Intelligent Transportation Systems **15** (2), pp. 494–506 (2014).

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.

J. F. P. Kooij, F. Flohr, E. A. I. Pool, and D. M. Gavrila, *Context-Based Path Prediction for Targets with Switching Dynamics,* Int. Journal of Computer Vision (IJCV) **127** (3), pp. 239–262 (2019).

N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, *DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents,* in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2017) pp. 336–345.

S. Lefèvre, D. Vasquez, and C. Laugier, *A survey on motion prediction and risk assessment for intelligent vehicles,* ROBOMECH Journal **1** (1), p. 1 (2014).

Y. Li, X.-Y. Lu, J. Wang, and K. Li, *Pedestrian Trajectory Prediction Combining Probabilistic Reasoning and Sequence Learning,* IEEE Trans. on Intelligent Vehicles , pp. 1–1 (2020).

S. Martin, S. Vora, K. Yuen, and M. M. Trivedi, *Dynamics of Driver's Gaze: Explorations in Behavior Modeling and Maneuver Prediction,* IEEE Trans. on Intelligent Vehicles **3** (2), pp. 141–150 (2018).

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.

S. Neogi, M. Hoy, K. Dang, H. Yu, and J. Dauwels, *Context Model for Pedestrian Intention Prediction Using Factored Latent-Dynamic Conditional Random Fields,* IEEE Trans. on Intelligent Transportation Systems , pp. 1–12 (2020).

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.

5

M. T. Phan, V. Fremont, I. Thouvenin, M. Sallak, and V. Cherfaoui, *Recognizing Driver Awareness of Pedestrian,* in *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)* (2014) pp. 1027–1032.

E. A. I. Pool, J. F. P. Kooij, and D. M. Gavrila, *Crafted versus Learned Representations in Predictive Models - A Case Study on Cyclist Path Prediction,* Submitted to IEEE Trans. on Intelligent Vehicles (2020).

E. A. Pool, J. F. Kooij, and D. M. Gavrila, *Using Road Topology to Improve Cyclist Path Prediction,* in *IEEE Intelligent Vehicles Symposium (IV)* (2017) pp. 289–296.

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.

A. Rangesh and M. M. Trivedi, *When Vehicles See Pedestrians With Phones: A Multicue Framework for Recognizing Phone-Based Activities of Pedestrians,* IEEE Trans. on Intelligent Vehicles **3** (2), pp. 218–227 (2018).

A. Rasouli, I. Kotseruba, and J. K. Tsotsos, *Understanding Pedestrian Behavior in Complex Traffic Scenes,* IEEE Trans. on Intelligent Vehicles **3** (1), pp. 61–70 (2018).

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.

D. A. Ridel, N. Deo, D. Wolf, and M. M. Trivedi, *Understanding Pedestrian-Vehicle Interactions with Vehicle Mounted Vision: An LSTM Model and Empirical Analysis,* in *IEEE Intelligent Vehicles Symposium (IV)* (2019) pp. 913–918.

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.

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.

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.

A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, *Human motion trajectory prediction: a survey,* The Int. Journal of Robotics Research **39** (8), pp. 895–935 (2020).

H. Saptoadi, *Suitable Deceleration Rates for Environmental Friendly City Driving,* Int. Journal of Research in Chemical, Metallurgical and Civil Engineering **4** (1), pp. 2–5 (2017).

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.

S. Siddharth and M. M. Trivedi, *On Assessing Driver Awareness of Situational Criticalities: Multi-modal Bio-Sensing and Vision-Based Analysis, Evaluations, and Insights,* Brain Sciences **10** (1), p. 46 (2020).

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.

J. Stapel, M. E. Hassnaoui, and R. Happee, *Measuring Driver Perception: Combining Eye-Tracking and Automated Road Scene Perception,* Human Factors (2020), 10.1177/0018720820959958.

SWOV, *Crossing facilities for cyclists and pedestrians,* SWOV Fact sheet , pp. 1–6 (2010).

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).

B. Völz, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, *Inferring Pedestrian Motions at Urban Crosswalks,* IEEE Trans. on Intelligent Transportation Systems **20** (2), pp. 544–555 (2019).

Y. Wang, Y. Ren, S. Elliott, and W. Zhang, *Enabling Courteous Vehicle Interactions through Game-based and Dynamics-aware Intent Inference,* IEEE Trans. on Intelligent Vehicles , pp. 1–1 (2019).

G. Welch and G. Bishop, *An Introduction to the Kalman Filter,* In Practice **7** (1), pp. 1–16 (2006).

WHO, *Global Status Report on Road Safety,* World Health Organization , pp. 1–20 (2018).

H. Winner, *Fundamentals of Collision Protection Systems,* in *Handbook of Driver Assistance Systems* (Springer, 2016) pp. 1149–1176.

C. Wu, Y. Lin, and W. Zhang, *Human Attention Modeling in a Human-Machine Interface based on the Incorporation of Contextual Features in a Bayesian Network,* in *IEEE Int. Conf. on Systems, Man and Cybernetics* (2005) pp. 760–766.

6

Discussion and Conclusions

This chapter summarizes the main findings from each chapter and discusses implications for the objectives and questions detailed in Chapter 1. Unanswered and newly identified questions are proposed for future work.

6.1. Effects of real-world automation on attentional resources

Chapter 2 set out to quantify how cognitive load differs between real-world manual and attentively supervised SAE2 automated driving, and compared this in a Tesla model S on two highways of different traffic complexity (a quiet, monotonic highway and a more engaging round-way of Amsterdam) and between automation-experienced Tesla owners and automation-inexperienced drivers. Changes in workload were monitored with the NASA RTLX, an auditory detection-response task (DRT) and cardiovascular activity. Additional ratings for driver trust and sleepiness were collected.

The subjective and secondary task measures for workload differed between driving conditions, but no interaction between automation use and environment emerged. This suggests that drivers remain sensitive to changes in task demand, and that automation does not affect this sensitivity.

When using the automation, automation-experienced drivers perceived a lower workload, while automation-inexperienced drivers perceived their workload to be similar to manual driving. Automation-experienced drivers also reported a higher trust in the automation compared to automation-inexperienced drivers. This indicates that a workload benefit is only perceived by automation-experienced users. It also suggests that no control is relinquished during first two hours of use before asserting the automation's basic capabilities. This experience and trust prerequisite for reduced perceived workload was not observed in simulator studies (de Winter *et al.*, 2014), but has an important implication. It suggests that early encounters with automation limitations are more likely to be resolved safely compared to later encounters. System capabilities and limitations may therefore better be demonstrated during the first drives rather than later, such that expectations reflect system's actual rather than desired capabilities, and automation surprises become less likely to arise later, as also demonstrated by Ebnali *et al.* (2019). This does not mean that early encounters of automation limitations are guaranteed to be safe. Victor *et al.* (2018) found that 21% of participating drivers allowed their automated vehicle to crash into a stationary object after 30 minutes of automated driving in a test track study despite attentive monitoring and having reasonable time to intervene (5 seconds or 14 seconds when noticing an early cue on the object's presence). Crash rates did not reduce when informing or demonstrating the possibility of automation failure prior or during the drive. Such demonstrations are therefore preferably provided in a safe format, such as with video, simulation or under guidance of a safety driver.

A range of experimental studies (de Winter *et al.*, 2014) reported that supervised automation gives a lower workload compared to manual driving. In these studies a reduced workload was found using both subjective rating on the NASA-TLX and rating scale mental effort, and using objective evaluation measuring performance on a non-driving task. Our study found a workload reduction only in experienced users, and only for their subjective RTLX ratings. Attentional demand as measured with the DRT suggests workload was higher while supervising automation for both participant groups. These observations indicate a miss-match between perceived and objective load. There can be several underlying reasons for this dissociation (Hancock and Matthews, 2018). Low driver demand from lane keeping and speed regulation and increased demand from monitoring extrinsic instead of intrinsic performance were proposed explanations for the objective workload increase. We also highlighted that the small sample size and incentives to monitor attentively prevent these findings to generalize to the full population for all driving conditions. However, this does not negate the observation that at least some users can obtain a higher average workload from supervising automation compared to manual driving, and it is thus possible for automation to be used as remedy against underload under engaging and monotonic traffic conditions. The results therefore demonstrate that attentive supervision of SAE2 can result in a healthier workload compared to manual driving, but the low perception of workload could discourage attentive monitoring if this perception is attributed to a low monitoring need. An implication is that the ironies of automation reside in our experiences and beliefs rather than being a limitation of physiology, and while these are by no means easy to alter, suggests it may be possible for automated driving to improve driver attention. The focus should therefore be directed to problems related to driver understanding and expectation of automated vehicles; Neuhuber *et al.* (2020) found in a 100 participant on-road study that within each age group, more than 50% of participants made at least one remark suggesting a fault in their mental model of the system, despite trusting the automation. Similarly, Farah *et al.* (2020) indicate that drivers often assume a larger operational design domain than what is described in the vehicle manual when it comes to automation capabilities.

Future work should test the generalizability of these workload findings, and consider calibrating workload perception and system limitation understanding rather than actual task demand when attempting to encourage attentive supervision. It should also be examined how much further automation can progress towards apparent autonomy before this potential benefit of strategic supervision on energetic state disappears. With SAE3 systems, the driver will have the choice to monitor or take the eyes off the road. It is unlikely that continuous monitoring of an SAE3 system will be as engaging as supervising today's SAE2 automation. On the other side of the spectrum, Scallen *et al.* (1995) found that alternating between 15 s manual driving and 15 s automation use raises workload and performance compared to longer periods, though users reported this transition rate to be distracting, and may not be desired for longer periods of use. There may be a sweet spot between SAE2 and SAE3 where reliability is consistent enough for the user to develop a clear understanding of the operational design domain, while the need for intervention remains frequent enough for strategic supervision to remain engaging and thus provide the intrinsic compulsion needed to stay attentive (Hancock, 2013). Lu *et al.* (2019) and Zhang *et al.* (2019b) demonstrate in a monitoring and takeover request study where automation dynamically transitioned between SAE3, SAE2 and manual driving, that the rate at which drivers motorically prepare to intervene while using SAE2 automation depends on how recently a need to intervene occurred, with more and earlier motoric preparation when such events occurred more recently. It should be examined how much further automation can progress towards apparent autonomy before this potential benefit of strategic supervision on energetic state disappears, and how the need for supervision can be maintained as its reliability progresses.

6.2. Naturalistic and longitudinal effects of automation

Chapter 3 explored how experience with automation changes automation usage and attention over the first two months of use in comparison to a one month manual baseline. We further presented our methodology and performance for two data enrichment methods on the naturalistic dataset.

Which conditions affect when automation is used?

Automation was mostly used on road types for which the systems are intended (motorways and highways), and use on urban roads was incidental rather than habitual, which suggests that users are aware of the system's general limitations and typically act accordingly. These observations of usage are in line with the distribution of activations across road types as found in a one month, 120 participants naturalistic study conducted in the Washington DC region (Russel *et al.*, 2018), but the suggested implication of users being aware of system limitations contradicts the findings from Farah *et al.* (2020) where a city road with unmarked road edges was correctly considered to be outside the automation's ODD of a Tesla model S by only 31% of their participants, which included frequent users of this system. This suggests that drivers may limit their SAE2 automation use to highway conditions even when convinced that urban settings are within the capabilities of the automation. Monitoring behaviour during moments of urban automation use and urban activation attempts could be inspected to investigate whether urban usage (attempts) are performed responsibly. Highway ACC+LK usage across driving speed suggests that users were generally comfortable using these systems during most highway traffic conditions, but did not use automation in brief periods of slow highway traffic. This distinction between short and longer periods of slow highway driving should be examined further.

No significant time in trip, time of day or experience effects were found for automation usage. While such effects may still exist, they at least are too subtle to emerge on the examined data from the plethora of factors which influence when automation is used. There is therefore no indication that automation is used more often at a particular time, and usage increases nor diminishes across the first two months of use, which suggests that drivers are generally willing to use automation regardless of time or experience. Whether automation use interacts beneficially or concernably with the circadian effects of performance remains to be investigated, as also emphasised by Kaduk *et al.* (2020).

How does automation use affect attention behaviour?

On the highway, head heading and pitch deviation during SAE2 automation use did not differ from manual baseline driving. No large differences in monitoring activity during SAE2 automation compared to manual driving were thus identified, which suggests that drivers generally remained attentive during automated driving, though more subtle differences in attention allocation cannot be ruled out. Both head heading and pitch deviation were however smaller during ACC use compared to baseline, and head heading deviation was larger for periods of manual driving in the experimental phase. Head pitch deviation also increased over the first 6 weeks of automation use for ACC+LK use and head heading deviation tended to increase during ACC and ACC+LK, which hints at behavioural adaptation. Further research is needed to assess if these changes in head deviation indicate better or worse monitoring behaviour. If drivers were mostly monitoring attentively during automation, increased attentional demand as found in Chapter 2 could be used to interpret lower head pose deviation (as found during ACC and initial ACC+LK use) as an increase in attention to road centre or cognitive narrowing due to an increased mental demand. However, it can also result from driving-unrelated tasks or thoughts (Victor *et al.*, 2005; Wang *et al.*, 2014), a reduced perceived need for visual scanning, or an increase in mind wandering (He *et al.*, 2011). Classification of driver attention to driving-related and -unrelated areas may provide additional insights. While automation of such classification was not successful in the present study, the identified changes in head pose deviation provide a motivation for further investigation, preferably using eye tracking rather than head pose.

The increase in head heading deviation from baseline-manual to experimental-manual highway driving may indicate that automation use can have an influence on attention patterns during manual driving. We deem it unlikely that the increase in head heading deviation from manual-baseline to manual-experimental indicates a carry-over effect, and this is more likely a consequence of strategic automation use. This hypothesis could be tested through selection of moments where strategic (dis)use is likely to emerge such as during lane changes as available from the Mobileye system, which records lane position and the location of other vehicles, or by testing if the higher head heading deviation also occurs for highway trips from the experimental condition where automation was not used at any point.

6.3. SUPPORTING DRIVER PERCEPTION OF INDIVIDUAL ROAD USERS

Automated driving technology provides a wealth of information about the situation surrounding the vehicle. This information can be used for additional purposes besides automated driving. One particular application addressed in this dissertation is its use in driver attention support. By combining information about surrounding road users with driver state monitoring, it may be possible to make specific and contextually relevant inferences about the driver's awareness of these road users. The conjecture is that specific support is better than generic support: being informed that your physiology or driving style indicates a general state of drowsiness gives little useful information, while being notified about a particular threat you overlooked can be more beneficial to safety. This specificity may also improve the acceptance of such support and can provide more opportunities to design for suitable levels of reliance and compliance by the driver. During manual driving, such support may identify attentional lapses or prevents attentional bias in expectation-defying scenarios, and whilst supervising automated driving, it may provide the performance feedback needed to remain attentive (Norman, 1981).

Chapters 2 and 3 investigated the human interaction with SAE2 automation in highway driving conditions. We now discuss Chapters 4 and 5 which investigated human perception in relation to other road users in urban conditions. In these studies participants drove manually and we investigated how the environment perception of the vehicle can support the driver by monitoring the driver's awareness towards individual road users during left turns (4) and using this cue to provide earlier predictions of collision risk in a crossing pedestrian scenario (5).

RECOGNITION TASK FOR EVALUATING AWARENESS TOWARD INDIVIDUAL ROAD USERS

Chapter 4 examines if driving automation technology can interpret the driver's awareness towards individual road users. Driver gaze is associated with surrounding road users as detected by computer vision during left turns on urban intersections. A post-drive recognition task was performed to assess driver awareness. In a logistic regression model, gaze behaviour could predict the relevance of road users, but not the driver's performance on the recognition task. The recognition task was sensitive to road user relevance and minimum gaze angle, and yielded a low false positive rate, which demonstrates it can identify awareness of individual road users during left turn manoeuvres. However, true positive rate was unexpectedly low; at best 40% for relevant attended road users. The low recognition rate is likely caused by memorization challenges. The 60 s delay, manoeuvre's demand on working memory, and visual encoding deficiencies were suggested as contributing factors which may be partially alleviated through design improvements. Since the recognition task remained sensitive despite low recall rate, it warrants further development. Potential improvements include reduction of time delay and intermittent tasks (e.g. finding parking location), and encoding more driving-relevant contextual cues in the recall task (e.g. motion, behaviours and spatial context). Following these improvements, more controlled evaluations in comparison to established SA measurement techniques will be required to evaluate the potential of the recognition task.

MONITORING SA FROM GAZE METRICS RELATIVE TO INDIVIDUAL ROAD USERS

We parameterized gaze behaviour relative to nearby road users and demonstrated that gaze duration eccentricity up to 10°, number of fixations and saccade angle were able to discriminate relevant from irrelevant road users with an accuracy of 73%. The gaze metrics could not predict the outcome of the recognition task. One explanation is that the forgetting aspect could not be captured by our model. While this prevents us from interpreting gaze patterns of unrecognized road users, the recognition task did provide useful insights. 18% of the road users that never entered the useful field of view (<10°) were still recognised, highlighting the importance of peripheral vision (Wolfe *et al.*, 2017). Hence, we strongly recommend that perception models incorporate more than fixation location in their parameterization.

To this day, no appropriate method exists to monitor the driver's awareness towards all individual road users in complex scenarios like left turns at urban intersections. While the recognition task's potential warrants further development, alternative approaches should be explored as well. One such direction is to see if Steady-State Visual Evoked Potential (SSVEP) could provide such method. SSVEP is a procedure where multiple flickering stimuli are presented, and the frequency response in the visual cortex is monitored to deduce which stimuli is attended. It has been used in brain research for over 50 years (Vialatte *et al.*, 2010) and has proven a reliable tool in fundamental attention research. The suggestion to use SSVEP for driver attention research is not new (Reddy *et al.*, 2007) but to my knowledge has never been attempted. If successful, this method may provide a reliable attention classifier for the perception phase of obtaining situation awareness.

Can awareness monitoring provide a prediction benefit for collision warning and intervention systems?

Chapter 5 evaluates if driver gaze and head pose eccentricity can provide a temporal advantage to emergency alerting or intervention systems. A crossing pedestrian collision risk prediction system is used as a case study to assess the benefits of measuring gaze and various other contextual cues. We evaluated the incremental benefits of six models with varying access to contextual cues on collision prediction between a vehicle and crossing pedestrian. By using a Dynamic Bayesian Network (DBN) structure, these models demonstrate how gaze and head pose observations influence awareness likelihood.

Contextual cues improved path predictions for scenarios with motion transitions of either the vehicle or the pedestrian while performing slightly worse on scenarios with unchanged dynamics. Introducing contextual cues about the pedestrian's awareness of the vehicle (head orientation), the position relative to the curb and the notion of a possible collision if both driver and pedestrian were to continue their present movement improved pedestrian path prediction as well as collision risk prediction. Similarly, contextual cues about a possible collision, the vehicle's distance to the pedestrian's intended path (*NCL*) and cues for driver awareness resulted in better vehicle path prediction, and better collision risk prediction for prediction horizons between 0.75 s and 2 s. Driver gaze did not provide path or collision risk prediction benefits over driver head pose, though gaze had better accuracy on the has-seen-pedestrian (*HSP*) context state compared to head pose. A performance advantage of gaze eccentricity over head-pose may emerge in more challenging scenarios when discriminating awareness towards multiple road users. Such evaluations would require natural as opposed to instructed viewing behaviour, and the presence of multiple spatial regions competing for attention.

6

It was also found that joint awareness (i.e. pedestrian being aware of vehicle and vehicle being aware of pedestrian) cannot always predict who will yield. In situations where both driver and pedestrian are aware of each other's presence, mutual awareness may be more discriminative as suggested by (Wang *et al.*, 2019) (i.e., estimating the other's awareness of oneself), and may benefit from cues on who has right of way.

The results also demonstrated that context-aware prediction models may suffer when outcomes defy the cue-based expectations. While probabilistic modelling such as provided by the used DBN framework can incorporate these exceptions in their likelihood estimates, emergency systems should avoid dependence on the correctness of these cues. Accidents after all tend to occur in situations which defy the expectations. Great care should thus be taken when selecting the desired likelihood ratio. Awareness monitoring can therefor aid traffic safety by providing a better likelihood estimate for collision risk, but will not provide certainty.

6.4. Implications Chapters 2 and 3: Is current SAE2 automation safe?

The findings from Chapters 2 and 3 portray a positive indication that SAE2 can be beneficial rather than detrimental to the driver's energetic state when used attentively and that experienced users seem to use these systems responsibly. Unfortunately there are a few nuances that warrant caution on the adoption of SAE2 automation. As already mentioned, the benefit of supervised automation only applies to attentive supervision, and providing an effective incentive to do so will be challenging. At present, some users of SAE2 automation indicate they engage in visual secondary tasks during periods of automated driving to prevent boredom (Lin *et al.*, 2018), and despite improvements through over-the-air updates, currently on-market automation seems to not support the driver sufficiently in adhering to the monitoring task (Banks *et al.*, 2018). Even in manual driving, a substantial amount of drivers engage in cognitively or even visually distracting activities, for example Oviedo-Trespalacios (2018) indicates that over 33% of interviewed Australian drivers engage daily in texting and browsing on their phone while driving. Furthermore, Young and Stanton (2007) already demonstrated that drivers respond later to to hazards when supervising SAE1 to SAE3 automation compared to manual driving. Despite several improvements through careful design of control transition aids, the response times remain high when using driving automation (Zhang *et al.*, 2019a), especially when these aids are not fully reliable (Zhang *et al.*, 2019b) as is the case for SAE2 automation, or when visually distracted with SAE2 automation (Louw *et al.*, 2019). It is for these reasons that views on whether SAE2 automated driving enhances safety are not unanimous (Dijsselbloem *et al.*, 2019; Hancock, 2019; de Winter, 2019).

These cautions indicate that while the results of Chapters 2 and 3 demonstrate that beneficial results from supervised automation are possible, the design space where these benefits are accomplished may be small. These benefits might also not be achieved for all types of drivers, since views towards automated driving are diverse (Kyriakidis *et al.*, 2015), and it is possible that the volunteering participants in Chapter 3 and Russel *et al.* (2018) represent a relatively safe category of drivers.

For the success of automated driving, it is imperative to find this design space. Our findings indicate that the desired safety benefit may already be available in on-market SAE2 automation and that it can be improved when attentive supervision is encouraged. This also indicates that there is virtue in continuing the development of effective strategies to improve driver attention during automated driving. The strategies reviewed and classified by Cabrall *et al.* (2019) either give the driver a more active role in the driving task or make it easier for the driver to stay engaged in the supervisory role. All strategies use the automation to collaborate with or support the driver, where also the methods and findings from chapters 4 and 5 may find application.

This also raises the question how automation technology should progress from SAE2 to SAE3 and SAE4, where the objective is to reduce engagement in the driving task. It seems that any intermediate levels between SAE2 and SAE3 should be avoided, both in design and in intended use. Ultimately the net traffic safety benefit intended by SAE2 and SAE3 can only be validated through crash statistics, which are slowly accumulating for SAE2 and SAE3 automation (Boggs *et al.*, 2020; Dijsselbloem *et al.*, 2019). However, much more data is needed before these statistics can test this safety benefit (Lindman *et al.*, 2017). In the meantime, surrogate safety metrics such as time to collision and driving safety field (Mullakkal-Babu *et al.*, 2017), and on-road studies such as presented in this dissertation may provide indications on how to guide automated driving to success.

6.5. IMPLICATIONS CHAPTERS 4 AND 5: CAN AUTOMATION TECHNOLOGY SUPPORT THE DRIVER'S MONITORING TASK?

Chapters 4 and 5 examine the approach of monitoring the driver's awareness to other road users and use it to support the driver; Chapter 4 inferred driver awareness from gaze metrics in relation to detected road users and focused on improving specificity of awareness monitoring by evaluating awareness towards individual road users. Chapter 5 implements such inference for making earlier predictions of intentions and behaviours.

In both chapters strategies were developed which may benefit each other. The deterministic regression model of Chapter 4 may be replaced by a probabilistic approach for estimating awareness likelihood as done in Chapter 5. Similarly, the driver-awareness aware models from Chapter 5 could benefit from the additional gaze metrics which were found to be predictive in Chapter 4, provided that these models are exposed to naturalistic pedestrian crossing encounters where multiple road users are competing for the driver's attention. Both studies may further benefit from a temporal (as opposed to per-trial) association between gaze and events, such that they can assess if and when the driver notices the appearance of new road users, changes in their behaviour, and the contextual cues which may precede such changes.

The findings also demonstrated a limitation of using contextual cues in driver support. While Chapter 5 showed that access to more contextual cues generally improved prediction performance, performance degraded for events which contradicted the learned rules. Similarly in Chapter 4, drivers occasionally were aware of road users despite never glancing near them. While using contextual cues improves the interpretation of what the driver is aware of and what road users may do in the near future, care should be taken to prevent that these cues (or attention-based architectures in general) introduce expectation bias to the automation's perception, at least not to the extent where the automation becomes susceptible to the same mistakes that human drivers tend to make. This insight forms an important distinction between the objectives for automated driving and driver support. While both applications will benefit from making fewer mistakes compared to human drivers, automated driving will find it easier to appease societal acceptance when the mistakes it makes exclude any that would have been easily avoided by a human driver Madhavan *et al.* (2006), whereas the faults by the driver and the support system ideally form disjoint sets. Ideally, driver support systems use models of human bias (such as top-down saliency maps, e.g. (Xia, 2019)) to predict when drivers are prone to overlook a relevant event, but use other methods to obtain their own road scene understanding. The difference between these three objectives is illustrated in figure 6.1. This does not mean that contextual cues should be avoided in obtaining system road scene understanding, but developers should be aware that the extent of its use may depend on the intended application.

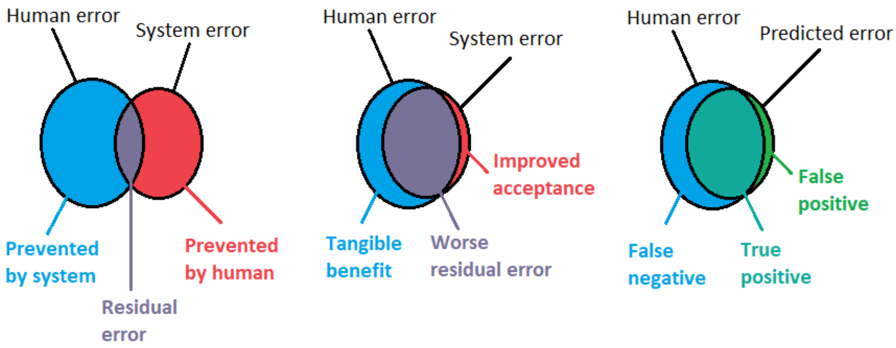


Figure 6.1: Illustration of three design objectives. Left: super-human system optimized to support the driver in avoiding errors. It prevents most human error but raises false alerts. Middle: a system with similar performance is optimized for automated driving, making fewer human-like errors compared to drivers, and avoiding non-human error. Right: A system attempting to predict human error.

We demonstrated that incorporating detailed scene perception as obtained by automated driving technology has the potential to infer whether the driver is aware of individual road users and their behaviour, though further research is needed for its development as well as how its performance can be evaluated with sufficient accuracy. If this path towards specificity gets further developed, it will provide a considerable benefit over traffic-agnostic monitoring techniques which can only indicate an aggregate level of attentiveness. Association of driver gaze to traffic cues should reasonably be expected to improve models of driver attention towards specific road users. The applications are not limited to supporting driver attention during manual driving; it may also provide driver support through early warning and intervention systems as explored in chapter 5, or even be used during supervised automation to re-introduce feedback on task performance (Norman, 1981) in SAE2 automation. The joint analysis of driver gaze and road scene can even benefit the development of automated driving through imitation learning as reviewed by Zhang *et al.* (2020) and used to identify driving scenarios where braking is required in Aksoy *et al.* (2020), or by learning computer vision to identify "action inducing" road users instead of all instances of a particular object (Xu *et al.*, 2020). But care should be taken that such systems do not become susceptible to the same mistakes as drivers tend to make. While promising data-driven (Fang *et al.*, 2019) and design-driven (Pal *et al.*, 2020) approaches to achieve this are being developed, validation will ultimately require a reliable way to distinguish between awareness and inattentional blindness or look but did not see phenomenon.

6.6. USING DRIVING AUTOMATION TECHNOLOGY FOR REAL-WORLD DRIVER RESEARCH

Chapter 1 argued that human factors research on automated driving should move from the simulator into the real world, since experience with real-world complications may benefit the development of systems which work as intended under these conditions. This dissertation pursued this goal by avoiding the simulator entirely and conducting all research on-road; for gaining such experience and to demonstrate that on-road research is becoming more practical and accessible than ever before, thanks to recent developments in automation. The pursuit of this demonstration led to both successes, testifying for the feasibility and importance of on-road driving research, and failures, which provide lessons on the importance for careful design of on-road experiments.

In Chapter 2, automation-inexperienced drivers did not perceive a lower workload, whereas simulator studies found such reduction for this participant group (de Winter *et al.*, 2014). We postulate that driver trust in the automation's ability to handle the complexities of real-world traffic (and the risk validity issue in a simulator) accounts for this difference.

In contrast, we were unable to infer any workload changes from cardiovascular activity; the derived measures (heart rate, sdNN, LF/HF) varied over time without apparent relation to the conditions or observed events, apart from a small decreasing trend in heart rate over time for automation-inexperienced drivers which may indicate acclimatization to the vehicle and automation. While this may partially be attributed to sensor quality, a reasonable conjecture is that heart rate is not specific enough to provide a reliable mental workload indicator unless great care is taken to control experimental conditions, in particular anything that involves physical effort and activities which alter breathing, such as speaking. We therefore recommend that cardiovascular activity is not used for the measurement of workload in semi-naturalistic studies, unless the confounding effects can be modelled out. It may however still have practical use in the measurement of other constructs such as driver drowsiness or anxiety (Kundinger *et al.*, 2020).

Working with the SAE2 naturalistic dataset in Chapter 3 has provided three important insights. Firstly, the initially unsuccessful retrieval of automation status from CAN data among four vehicle types and the additional salvaging efforts that were needed to retrieve (some of) this information demonstrate the importance of careful design, but also validation of the data acquisition system. In on-road driving, no information comes for free. Secondly, the automation status retrieval for the Tesla from image recognition demonstrates that even very simple implementations of machine learning can make the difference between success and failure of data retrieval. With only 17 neurons, a simple neural network improved classification accuracy from 70% obtained from logistic regression to 99.33%, while the use of template matching prevented the need for training of (and the required annotation for) a complete convolutional neural network for icon detection. Thirdly, a similar enrichment attempt was unable to reliably classify attended regions from head pose, demonstrating that head pose without information of gaze direction may be insufficient for region of attention classification. This provides one demonstration (among many) that machine learning can only be as informative as the data on which it is trained. While continued research in using (possibly cheaper and more robustly obtained) head pose for attention inference is to be encouraged, using head pose as an attention measurement tool without inclusion of gaze data is not recommend for research purposes, and this requires careful instrumentation and processing design.

Despite these successes and experiences, it cannot be denied that on-road research requires considerable effort in its preparation to be successful, and that the work and knowledge required may be prohibitive for many researchers to which the instrumentation forms the means rather than the end. The key solution is interdisciplinary collaboration on a single research platform and maintaining it across multiple generations of researchers to build upon each other's work. Within the Intelligent Vehicles group at the TU Delft, this is realized on an instrumented Prius (Ferranti *et al.*, 2019) through the contributions of several engineers, researchers, students and collaborating organizations (figure 6.2). The maintainability is achieved through platforms such as the Robot Operating System (ROS) and GIT, technical staff and incorporating maintainability practices in education. These efforts have made it possible to extend its use from a platform for vehicle automation research to one for driver support research, and from a research platform to a classroom for education, where bachelor and master students develop driver support systems and experience the practical challenges of field operational testing first-hand. Additionally, a considerable instrumentation effort can be avoided when manufacturers collaborate in such research by supporting access to CAN data and other information readily available in the vehicle.

Figure 6.2: The vehicle, and the team (October 2018).

6.7. CONCLUSIONS

CHAPTER 2: EFFECT OF SAE2 AUTOMATION ON DRIVER WORKLOAD.

Drivers remain sensitive to changes in task demand while supervising automated driving. In contrast to expectation, SAE2 automation raises workload when monitored attentively. This can be beneficial for driver attention, but perception of workload during supervision may be too low for this to occur naturally. Future work should test the generalizability of these workload findings, and consider calibrating workload perception and system limitation understanding instead of calibrating the actual task demand when attempting to encourage attentive supervision. It should also be examined how much further automation can progress towards apparent autonomy before this potential benefit of strategic supervision on energetic state disappears.

Chapter 3: naturalistic automation usage and head activity

Automation is mostly used on road types where its use is generally considered suitable, with only incidental use on urban roads, which suggests that users are adhering to the operational design domain of these vehicles. On highways, automation is used at all speed conditions, but less during short periods of slow driving. No time-in-drive, time-of-day or effects of experience were found for automation use.

During highway automation use, head pose standard deviation did not differ between SAE2 automation and baseline manual driving, but head pitch deviation increased over the first six weeks of use, which hints at behavioural adaptation. Head heading and pitch deviation were smallest during ACC use. Further research is needed to assess if these differences indicate better or worse monitoring behaviour.

Chapter 4: measuring driver awareness

The recognition task is sensitive to both road user relevance and gaze behaviour, but could not be predicted by gaze metrics and requires further development to reduce forget rates. Further analysis is needed to compare the recognition task to established situation awareness measures after these improvements are obtained.

Relative gaze metrics such as eccentricity and number of saccades to an individual road user could predict road user relevance. Temporal association to events and contextual cues are recommended to improve the specificity of noticing relevant changes in road user behaviour. At least 18% of road users were recognised while only observed peripherally, suggesting that peripheral vision should not be neglected in attention monitoring.

Chapter 5: predicting collision risk with pedestrian

In addition to other contextual cues, driver and pedestrian attention monitoring can provide a better prediction of collision risk with a crossing pedestrian when predicting further than 0.75 seconds ahead. By using a DBN structure, these models incorporated how gaze and head pose observations influence awareness likelihood. Context-aware prediction models perform worse when scenario outcomes defy the cue-based expectations.

6.8. References

E. Aksoy, A. Yazıcı, and M. Kasap, *See, attend and brake: An attention-based saliency map prediction model for end-to-end driving,* arXiv (2020).

V. A. Banks, A. Eriksson, J. O'Donoghue, and N. A. Stanton, *Is partially automated driving a bad idea? observations from an on-road study,* Applied ergonomics **68**, pp. 138–145 (2018).

A. M. Boggs, B. Wali, and A. J. Khattak, *Exploratory analysis of automated vehicle crashes in california: A text analytics & hierarchical bayesian heterogeneity-based approach,* Accident; analysis and prevention **135**, p. 105354 (2020).

C. D. D. Cabrall, A. Eriksson, F. Dreger, R. Happee, and J. de Winter, *How to keep drivers engaged while supervising driving automation? a literature survey and categorisation of six solution areas,* Theoretical Issues in Ergonomics Science **20** (3), pp. 332–365 (2019).

J. Dijsselbloem, M. van Asselt, S. Zouridis, and C. Verheij, *Wie stuurt? verkeersveiligheid en automatisering in het wegverkeer,* (2019).

M. Ebnali, K. Hulme, A. Ebnali-Heidari, and A. Mazloumi, *How does training effect users' attitudes and skills needed for highly automated driving?* Transportation Research Part F: Traffic Psychology and Behaviour **66**, pp. 184–195 (2019).

J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. & Li, *Dada-2000: Can driving accident be predicted by driver attention? analyzed by a benchmark,* IEEE Intelligent Transportation Systems Conference (ITSC) , pp. 4303–4309 (2019).

H. Farah, S. Bhusari, P. van Gent, F. A. M. Babu, P. Morsink, R. Happee, and B. van Arem, *An empirical analysis to assess the operational design domain of lane keeping system equipped vehicles combining objective and subjective risk measures,* IEEE Transactions on Intelligent Transportation Systems , pp. 1–10 (2020).

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,* IEEE Intelligent Vehicles Symposium , pp. 1660–1666 (2019).

P. A. Hancock, *In search of vigilance: The problem of iatrogenically created psychological phenomena,* American Psychologist **68** (2), pp. 97–109 (2013).

P. A. Hancock, *Some pitfalls in the promises of automated and autonomous vehicles,* Ergonomics **62** (4), pp. 479–495 (2019).

P. A. Hancock and G. Matthews, *Workload and performance: Associations, insensitivities, and dissociations,* Human factors **61** (3), pp. 374–392 (2018).

J. He, E. Becic, Y.-C. Lee, and J. S. McCarley, *Mind wandering behind the wheel: Performance and oculomotor correlates,* Human factors **53** (1), pp. 13–21 (2011).

S. I. Kaduk, A. P. J. Roberts, and N. A. Stanton, *The manual shift in phase: The impact of circadian phase on semi-autonomous driving. what can we learn from current understanding in manual driving?* Theoretical Issues in Ergonomics Science **20**, pp. 1–21 (2020).

T. Kundinger, A. Riener, N. Sofra, and K. Weigl, *Driver drowsiness in automated and manual driving: Insights from a test track study,* 25th International Conference on Intelligent User Interfaces , pp. 369–379 (2020).

M. Kyriakidis, R. Happee, and J. C. F. de Winter, *Public opinion on automated driving: Results of an international questionnaire among 5000 respondents,* Transportation Research Part F: Traffic Psychology and Behaviour **32**, pp. 127–140 (2015).

R. Lin, L. Ma, and W. Zhang, *An interview study exploring tesla drivers' behavioural adaptation,* Applied ergonomics **72**, pp. 37–47 (2018).

M. Lindman, I. Isaksson-Hellman, and J. Strandroth, *Basic numbers needed to understand the traffic safety effect of automated cars,* IRCOBI Conference (2017).

6

T. Louw, J. Kuo, R. Romano, V. Radhakrishnan, M. G. Lenné, and N. Merat, *Engaging in ndrts affects drivers' responses and glance patterns after silent automation failures,* Transportation Research Part F: Traffic Psychology and Behaviour **62**, pp. 870–882 (2019).

Z. Lu, B. Zhang, A. Feldhütter, R. Happee, M. Martens, and J. de Winter, *Beyond mere take-over requests: The effects of monitoring requests on driver attention, take-over performance, and acceptance,* Transportation Research Part F: Traffic Psychology and Behaviour **63**, pp. 22–37 (2019).

P. Madhavan, D. A. Wiegmann, and F. C. Lacson, *Automation failures on tasks easily performed by operators undermine trust in automated aids,* Human factors **48** (2), pp. 241–256 (2006).

F. A. Mullakkal-Babu, M. Wang, H. Farah, B. van Arem, and R. Happee, *Comparative assessment of safety indicators for vehicle trajectories on highways,* Transportation research record **2659** (1), pp. 127–136 (2017).

N. Neuhuber, G. Lechner, T. E. Kalayci, A. Stocker, and B. Kubicek, *Age-related differences in the interaction with advanced driver assistance systems - a field study,* International Conference on Human-Computer Interaction **22**, pp. 363–378 (2020).

J. C. F. de Winter, *Pitfalls of automation: A faulty narrative?* Ergonomics **62** (4), pp. 505–508 (2019).

J. C. F. de Winter, R. Happee, M. H. Martens, and N. A. Stanton, *Effects of adaptive cruise control and highly automated driving on workload and situation awareness, a review of the empirical evidence,* Transportation Research Part F: Traffic Psychology and Behaviour **27** (B), pp. 196–217 (2014).

D. A. Norman, *Categorization of action slips,* Psychological Review **88** (1), pp. 1–15 (1981).

O. Oviedo-Trespalacios, *Getting away with texting: Behavioural adaptation of drivers engaging in visual-manual tasks while driving,* Transportation Research Part A: Policy and Practice **116**, pp. 112–121 (2018).

A. Pal, S. Mondal, and H. I. Christensen, *Looking at the right stuff - guided semantic-gaze for autonomous driving,* IEEE Winter Conference on Applications of Computer Vision , pp. 11883–11892 (2020).

B. S. Reddy, O. A. Basir, and S. J. Leat, *Estimation of driver attention using visually evoked potentials,* Proceedings of the IEEE Intelligent Vehicles Symposium (2007), 10.1109/IVS.2007.4290179.

S. M. Russel, M. Blanco, J. Atwood, W. A. Schaudt, V. L. Fitchett, and S. Tidwell, *Naturalistic study of level 2 driving automation functions,* National Highway Traffic Safety Administration, DOT HS 812 642 (2018).

S. F. Scallen, P. A. Hancock, and J. A. Duley, *Pilot performance and preference for short cycles of automation in adaptive function allocation,* Applied ergonomics **26** (6), pp. 397–403 (1995).

F.-B. Vialatte, M. Maurice, J. Dauwels, and A. Cichocki, *Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives,* Progress in neurobiology **90** (4), pp. 418–438 (2010).

T. W. Victor, J. L. Harbluk, and J. A. Engström, *Sensitivity of eye-movement measures to in-vehicle task difficulty,* Transportation Research Part F: Traffic Psychology and Behaviour **8** (2), pp. 167–190 (2005).

T. W. Victor, E. Tivesten, P. Gustavsson, J. Johansson, F. Sangberg, and M. Ljung Aust, *Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel,* Human factors **60** (8), p. 1095–1116 (2018).

Y. Wang, B. Reimer, J. Dobres, and B. Mehler, *The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand,* Transportation Research Part F: Traffic Psychology and Behaviour **26**, pp. 227–237 (2014).

Y. Wang, Y. Ren, S. Elliott, and W. Zhang, *Enabling Courteous Vehicle Interactions through Game-based and Dynamics-aware Intent Inference,* IEEE Transactions on Intelligent Vehicles **5** (2), pp. 217–228 (2019).

B. Wolfe, J. Dobres, R. Rosenholtz, and B. Reimer, *More than the useful field: Considering peripheral vision in driving,* Applied ergonomics **65**, pp. 316–325 (2017).

Y. Xia, *Driver Eye Movements and the Application in Autonomous Driving,* Doctoral thesis, University of California, Berkeley (2019).

Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, *Explainable object-induced action decision for autonomous vehicles,* IEEE Winter Conference on Applications of Computer Vision , pp. 9523–9532 (2020).

M. S. Young and N. A. Stanton, *Back to the future: Brake reaction times for manual and automated vehicles,* Ergonomics **50** (1), pp. 46–58 (2007).

B. Zhang, J. de Winter, S. Varotto, R. Happee, and M. Martens, *Determinants of take-over time from automated driving: A meta-analysis of 129 studies,* Transportation Research Part F: Traffic Psychology and Behaviour **64**, pp. 285–307 (2019a).

B. Zhang, Z. Lu, R. Happee, J. C. F. de Winter, and M. H. Martens, *Compliance with monitoring requests, biomechanical readiness, and take-over performance: Video analysis from a simulator study,* ITS European Congress **13** (2019b).

R. Zhang, A. Saran, B. Liu, Y. Zhu, S. Guo, S. Niekum, D. Ballard, and M. Hayhoe, *Human gaze assisted artificial intelligence: A review,* International Joint Conference on Artificial Intelligence **29**, pp. 4951–4958 (2020).

Acknowledgements

On my journey towards this dissertation, I have received a tremendous amount of help and encouragement from many people.

First, I would like to thank my supervisors. Riender Happee, you have been an excellent guide. Thank you for giving me this opportunity to pursue my PhD, and for realizing and managing IAVTRM as well as all the adjacent projects; they provided an amazing environment in which to grow as a researcher. Thank you for your inspiration, trust, guidance, patience and encouragement along the way, and thank you for introducing me to so many people; both in industry and academia. Dariu Gavrila, thank you for encouraging me to be ever self-exceeding in my pursuit of scientific rigour, and thank you for leading the Intelligent Vehicles group to where it is today; many aspects of my research would not have been realized if it weren't for the team, tools and goals you have assembled.

I would also like to thank the members of the HFAuto project, who gave me the best introduction into the field of automotive human factors I could possibly wish for. Thank you Neville Stanton for sparking my interest in driver workload. Ignacio Solis, Alberto Morando, Daniel Heikoop and Alexander Eriksson, thank you all for extensively discussing and exchanging views on driver workload. Bo Zhang, thank you also for our discussions on driver workload and transitions of control. Thank you Silvia Varotto for confronting me with the right questions, and for your help at SWOV. Thierry Bellet and especially Matt Sassman, thank you both for turning my visit of Lyon and IFFSTAR into an experience I will never forget. Katja Kircher, Jan Andersson, Marco Dozza, Veronika Petrovych thank you for doing the same in Göteborg, VTI and Chalmers University.

I also want to thank my colleagues and fellow PhDs in Delft and elsewhere. Markus Roth, thank you for our intense collaboration, your visits and hospitality, and thank you for showing me that brewing a good coffee requires no less than two PID controllers. Joost de Winter, thank you for all your feedback on my project proposals and for your suggestions for relevant literature. Thank you also for all the cool student projects we conducted together, and sorry for not yet upgrading them to publications. David Abbink, thank you for your aspiring enthusiasm and for giving a crash course on human factors research. Thank you also for sharing a few insights on what to consider while writing a grant proposal. Sina Nordhoff, thank you for teaching me more about the analysis of interviews and thank you for our collaboration in on-going projects. Zhenji Lu, thank you for sharing the challenges of practicing human factors research while having an engineering background, and thank you for cautioning me for the difficulties of measuring situation awareness. Christopher Cabrall, thank you for your enthusiasm, your views on vigilance and driver attention, and thank you for helping me realize the virtue of not making experiments more complicated than necessary. Felix Dreger, thank you for guiding my thoughts with your knowledge on cognitive psychology and patient advice on the selection of statistical tests, and thank you for showing me all the highlights of Berlin. Pavlo Bazilinskyy, thank you for our collaborations with the Bachelor projects, and thank you for organizing many of our social events. András Pálffy, Thomas Hehn and especially Ewoud Pool, thank you for your help with ROS, git, machine learning and DBNs, and thank you for teaching me how to brew a decent beer, and how to ruin it with apple juice. Tugrul Irmak, thank you for your wits, your company and for helping me solve difficult problems, especially the ones in Fontainebleau. Natalia Kovacsova and Sarah Barendswaard, thank you for occasionally dragging everyone out off the office and into a cafe or to the pub. George Dialynas, Carlos Celemin, Sarvesh Kolekar, thank you too for providing the necessary distractions from work. Paul van Gent, thank you for our discussions on instrumentation and physiological measurements during our coffee breaks and thank you for your help with monitoring heart rate. Freddy Mullakal Babu, thank you for helping me with the execution of my first on-road experiment. Joris Domhof, thank you for helping me to improve the calibration of the eye tracker. Markus Braun, thank you for providing better pedestrian head pose detection. Yanggu Zheng, thank you for your feedback on my discussion section. Tom Dalhuisen, Frank Everdij and Ronald Ensing, thank you for turning the Prius into a maintainable research platform. Thank you also for improving my soldering skills and for encouraging me to write maintainable code, and for introducing me to the more advanced chapters of version control.

I also thank Marieke Martens, Marika Hoedemaeker, Nicole van Nes and Peter van der Knaap for arranging the possibility for me to contribute to the SAE2 naturalistic dataset. Michiel Christoph and Reinier Jansen, thank you for the crash course on SQL, and for the advice for improving my annotation methodology.

I also thank a number of students I supervised, who helped out exploring ideas and concepts during my research. Mounir El Hassnaoui, thank you for your help in designing and executing the perception monitoring experiment described in Chapter 4. Vishal Onkhar, thank you for your continued search for better ways to classify pedestrian and driver mutual awareness. Chris Smit, thank you for introducing me to decision tree classification and to the measurement of pupilometry. Thomas de Boer, Jim Hoogmoed, Nathan Looye, Jim van der Toorn and Rik de Vos, thank you for quantifying the accuracy of calibrating and compensating head mounted eye tracking using feature matching. Ruben Mabesoone, Robert Baelde, Casper Spronk, Paul Vroegindeweij, thank you for your contributions to both the hardware and software of the remote eye tracker. Coert de Koning, Hidde Lingmont, Tjebbe Lint and Florian van den Ouden, thank you for exploring a variety of auditory alerting concepts. Also thank you Frank Anema, for providing the test vehicle and other tools used in this project. Hugo Boer, Max Bottemanne, Geert van den Broek, Stijn Mijnster and Wouter van der Wal, thank you for uncovering the challenges of latency, tracking and user understanding in your crossing pedestrian alerting system. Gitte Hornung, Daan Koetzier, Tessa Talsma and Bob van der Windt, thank you for exploring the challenges of providing continuous attention feedback, and for your novel approaches to tackle them.

I am also greatly thankful to all my friends and family, who complemented my personal life, and without whom I would not have been able to complete this endeavour. Especially thanks to Jerom, Victor, Noud, Rudi, Ayla, Tjeu, Robbert, Robert and Floor for helping me find distraction and energy through sports, game and food; as well as to my housemates Willem, Peter, Roy, Jip and my brother Wout, whose company was always appreciated, but especially during the Covid-19 lockdown. I also want to express my deepest gratitude to my parents. Thank you for your all your love, wisdom and unconditional support throughout my life. Thank you also for reminding me to take enough rest, and for encouraging this through the many chores, hikes, sailing trips and visits.

Curriculum Vitæ

Jaap Cornelis Jork Stapel

14-07-1989 Born in Haarlem, the Netherlands.

Education

2008–2011	BSc Aerospace Engineering, Minor: Challenging Materials, Delft University of Technology, the Netherlands
2012–2015	MSc Aerospace Engineering: Control & Simulations, Delft University of Technology, the Netherlands. Electives include advanced control and robotics courses. Thesis on developing efficient estimation methods for dynamic flight envelopes using nonlinear partial differential equations.
2016–2020	PhD: Cognitive Robotics, Delft University of Technology, the Netherlands *Thesis:* On-Road Assessment of Driver Workload and Awareness in Automated Vehicles
2020–Present	Researcher (Post-doc): studying perceived safety in automated driving, Delft University of Technology, the Netherlands

Professional Experience

2012–2013 3 mo. internship	Boeing Research & Technology, Europe Madrid, Spain. Developed fault-tree analysis tools in the safety department.
2010–2015 dream team	Delft Aerospace Rocket Engineering Delft, the Netherlands. Designed, built and launched several small-scale rockets. Team leader capsule and recovery team during STRATOS II launch campaign.

List of Publications

Doctoral articles

The works of Chapter 2 have appeared in publication as follows:

J. Stapel, F.A. Mullakkal-Babu, R. Happee, *Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving*, Transportation research part F: traffic psychology and behaviour, **60** (2019).

J. Stapel, F.A. Mullakkal-Babu, R. Happee, *Driver behavior and workload in an on-road automated vehicle*, Proceedings of the RSS2017 Conference (2017).

The candidate was solely responsible for the production of the content with named co-authors providing support through review and modification only. The candidate contributed substantially to the conception, design, execution and analysis, and wrote the article. F.A. Mullakkal-Babu (F.M-B) provided support during data collection and reviewed the manuscript. Riender Happee (R.H) contributed to the conception and design of the study and reviewed the manuscript.

The works of Chapter 3 will be submitted as follows:

J. Stapel, R. Happee, M. Christoph, N. van Nes, M. Martens , *Exploration of the impact of SAE2 automation on driving behaviour: A naturalistic driving study* (unpublished).

The candidate was solely responsible for the production of the content with named co-authors providing support through review and modification only. The candidate contributed substantially to the conception, design, execution and analysis of the results presented. M. Martens (M.M) and N. van Nes (N.v.N) were involved in the conception, design, execution of the NDS study and database verification and preprocessing. M. Christoph (M.C) provided support in the analysis of this work. R.H and M.M contributed to the conception of the study. R.H, N.v.N and M.M contributed to the design of the study and reviewed the article.

The works of Chapter 4 have appeared in publication as follows:

J. Stapel, M. El Hassnaoui, R. Happee, *Measuring Driver Perception: Combining Eye-tracking and Automated Road Scene Perception*, Human Factors, (2020).

The candidate was solely responsible for the production of the content with named co-authors providing support through review and modification only. The candidate contributed substantially to the conception, design, execution, analysis, and wrote the article. M. El Hassnaoui (M.E.H) contributed to the conception, design, execution and analysis and contributed to the article. R.H contributed to the conception and design of the study and reviewed the manuscript.

The works of Chapter 5 is submitted as:

M. Roth, J. Stapel, R. Happee, D.M. Gavrila, This chapter has been submitted as: Markus Roth, Jork Stapel, Riender Happee, Dariu M. Gavrila, *Driver and Pedestrian Mutual Awareness for Path Prediction and Collision Risk Estimation*.

This work builds upon earlier work by M. Roth, F. Flohr and D.M. Gavrila, titled *Driver and Pedestrian Awareness-based Collision Risk Analysis*, which appeared at the IEEE Intelligent Vehicles Symposium 2016. In that conference paper, the basic concepts of this work were introduced. In this work (i.e. a journal upgrade), methods were enhanced, new and more extensive experiments were performed (incl. measuring driver head pose and gaze) and a more complete discussion was contributed. The candidate co-produced the content with M. Roth (M.R), with the other named co-authors providing support through review and modification only. The candidate contributed substantially to the design, execution, analysis, and writing of the article. M.R contributed substantially to the conception, design, execution, analysis, and writing of the article. D.M. Gavrila (D.G) contributed to the conception and the writing of the article. R.H and D.G contributed to the design of the study and reviewed the manuscript.

The individual contributions can be further detailed as follows. The proposed model architecture was conceived and developed by M.R and later refined collaboratively with the extension of the *near-crossing-line* node, the introduction of model variants and the procedure for dynamically updating model nodes during predictions. The instrumentation (sensor layout, data integration, synchronization and calibration) was designed by M.R and collaboratively refined. They were installed in the vehicle by the candidate. Integration of the eye tracker and alterations to the calibration procedure were provided by the candidate. The experimental procedures were developed by M.R and later refined collaboratively. Experiments were conducted collaboratively. Collected data was pre-processed by M.R. Existing code was adapted, extended and tested in close collaboration for mode training and evaluation. Decisions on trained and fixed parameters were made collaboratively. Values for initial and fixed parameters were developed by the candidate. Model training and evaluation was performed by M.R. M.R contributed substantially to the path prediction performance results and collision risk evaluation metric. The candidate contributed substantially to the evaluation of collision risk prediction and evaluation of driver attention state. The candidate and M.R contributed to the writing in close collaboration.

JOURNALS

1. **J. Stapel**, M. El Hassnaoui, R. Happee, *Measuring Driver Perception: Combining Eye-tracking and Automated Road Scene Perception*, Human Factors, (2020).

2. S. Nordhoff, **J. Stapel**, B. van Arem, R. Happee, *Passenger opinions of the perceived safety and interaction with automated shuttles: A test ride study with 'hidden' safety steward*, Transportation Research Part A: Policy and Practice, 138 (2020)

3. C.D.D. Cabrall, **J.C.J. Stapel**, R. Happee, J.C.F. de Winter, *Redesigning today's driving automation toward adaptive backup control with context-based and invisible interfaces*, Human Factors **2**, 62 (2020)

4. **J. Stapel**, F.A. Mullakkal-Babu, R. Happee, *Automated driving reduces perceived workload, but*

monitoring causes higher cognitive load than manual driving, Transportation research part F: traffic psychology and behaviour, **60** (2019)

CONFERENCES

1. **J.C.J. Stapel**, C.C. de Visser, E. van Kampen, Q.P. Chu, *Efficient methods for flight envelope estimation through reachability analysis*, AIAA guidance, navigation, and control conference, 83 (2016).

2. **J. Stapel**, F.A. Mullakkal-Babu, R. Happee, *Driver behavior and workload in an on-road automated vehicle*, Proceedings of the RSS2017 Conference (2017). (preliminary version of Chapter 2)

3. P. Bazilinskyy, **J. Stapel**, C. de Koning, H. Lingmont, T.S. de Lint, T.C. van der Sijs, F.C. van den Ouden, F. Anema, J.C.F. de Winter, *Graded auditory feedback based on headway: An on-road pilot study*, Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference: Varieties of interaction: from User Experience to Neuroergonomics (2018).

4. C. Cabrall, **J. Stapel**, P. Besemer, K. Jongbloed, M. Knipscheer, B. Lottman, P. Oomkens, N. Rutten, *Plausibility of Human Remote Driving: Human-Centered Experiments from the Point of View of Teledrivers and Telepassengers*, Proceedings of the Human Factors and Ergonomics Society Annual Meeting **1**, 63 (2019).

5. C.O. Smit, J.C.F. de Winter, **J. Stapel**, F. Doubek N. von Janczweski, *Estimating cognitive load with varying light intensity*, submitted to ACM CHI Conference on Human Factors in Computing Systems, (2020)

SUPERVISED MASTER THESES

1. C.O. Smit, *Estimating cognitive load with varying light intensity* (2020, co-supervised with F. Doubek, N. von Janczewski and J.C.F. de Winter)

2. J. Hoogmoed, *Interactions between driver, vehicle, and vulnerable road users* (in-progress, co-supervised with P. Bazilinskyy and J.C.F. de Winter)

3. V. Onkhar, *Algorithmic detection of eye contact in driver-pedestrian interactions* (2020, Co-supervised with P. Bazilinskyy and J.C.F. de Winter)

4. M. El Hassnaoui, *Measuring driver perception during on-road eye-tracking: Combining gaze behaviour and vehicle's road scene perception* (2019, Co-supervised with R. Happee. Leading to Chapter 4 in this thesis)

SUPERVISED BACHELOR STUDENT PROJECTS

Reports can be provided on request.

1. M. Knipscheer, B. Lottman, P. Oomkens, J. Schol, *How Is Driving Performance Impaired by Providing All Driving Visuals Through a Display?* (Bachelor end project 2016, co-supervised with C.D.D. Cabrall).

2. B. van Dijk, W.-J. Littel, R. Mak, D. Oudmaijer, *HMI-supported merging: driving near truck platoons* (Bachelor end project 2017, co-supervised with F. Dreger and J.C.F. de Winter)

3. C. de Koning, H. Lingmont, T.S. de Lint, T.C. van der Sijs, F.C. van den Ouden, *Adaptive real-time auditory feedback systems on headway* (Bachelor end project 2017, co-supervised with P.

Bazilinskyy, J.C.F. de Winter and F. Anema)

4. T.A.B de Boer, J. Hoogmoed, N.M. Looye, J.R.P. van der Toorn, R.P. de Vos, *Combining eye tracking with semantic scene labeling in cars* (Bachelor end project 2017, co-supervised with J.C.F. de Winter)

5. R.P.H. Mabesoone, R.E.W. Baelde, C.J. Spronk, P. Vroegindeweij, *Collaborative monitoring in an instrumented vehicle*, (Bachelor end project 2018, co-supervised with P. Bazilinskyy, F. Dreger and R. Happee)

6. S.L.Y. van Beuren, T.H.M. van der Laan, S.J. Uitendaal, J. van der Zeeuw, *Changing perspectives 2: getting the truck on the road by evaluating drive*, (Bachelor end project 2018, co-supervised with F. Dreger, R. Happee, J.C.F. de Winter)

7. B. ter Borg, L. Foorthuis, J. Tas, T. van Zee, *A Study into the Scenarios where Motorized Vehicles and Vulnerable Road Users Communicate*, (Bachelor end project 2018, co-supervised with F. Dreger and J.C.F. de Winter)

8. H. Boer, M. Bottemanne, G. van den Broek, S. Mijnster, W. van der Wal, *Enhancing pedestrian crossing alerting systems*, (Bachelor end project 2019, co-supervised with J.C.F. de Winter and R. Happee)

9. G. Hornung, D. Koetzier, T. Talsma, B. van der Windt, *AVA, a prototype for speech-based assistance in automated driving*, (Bachelor end project 2019, co-supervised with J.C.F. de Winter and R. Happee)

accompanying the dissertation

ON-ROAD ASSESSMENT OF DRIVER WORKLOAD AND AWARENESS IN AUTOMATED VEHICLES

by

Jaap Cornelis Jork STAPEL

1. Supervising SAE2 automation can prevent drowsiness and passive fatigue.
   This proposition pertains to Chapter 2.
2. The main obstacle for better attention support is the inability to measure driver awareness.
   This proposition pertains to Chapter 4.
3. The lack of specificity makes cardiovascular metrics useless for naturalistic mental workload monitoring.
   This proposition pertains to Chapter 2.
4. Supervised driving automation should be artificially limited.
   This proposition pertains to Chapter 6.
5. When training computer vision for road scene interpretation, different objective functions are needed for application in automated driving and driver support.
6. Promovendi need more guidance in placing expectations between aspiration and reality.
7. When plotting predictions against time, predictions should be drawn at the time for which they were made, not at the time when they were made.
8. Careful design of data collection is more important than careful design of tests or experimental conditions.
9. While the tower of science is made of ivory, the scaffolding is often not, and does not always need to be to fulfill it's purpose.
10. Writing a paper is not the process of transferring new knowledge, but the process of creating such knowledge.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotors Dr. ir. R. Happee and prof. dr. D.M. Gavrila.