From sight to insight: Eye contact and eye-tracking in the driver-pedestrian context

Onkhar, V.

DOI
10.4233/uuid:a68e8dd0-c3df-403b-a6f5-008c7d016676

Publication date
2025

Document Version
Final published version

Citation (APA)
Onkhar, V. (2025). *From sight to insight: Eye contact and eye-tracking in the driver-pedestrian context*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:a68e8dd0-c3df-403b-a6f5-008c7d016676

From sight to insight: Eye contact and eye-tracking in the driver-pedestrian context

Vishal Onkhar

Delft University of Technology

From sight to insight: Eye contact and eye-tracking in the driver-pedestrian context

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
Monday, 10 March 2025 at 15:00 o'clock

by

Vishal ONKHAR

Master of Science in Mechanical Engineering,
Delft University of Technology, The Netherlands
born in Chennai, India

This dissertation has been approved by the promotors:

Prof. dr. ir. J. C. F. de Winter
Dr. D. Dodou

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. ir. J. C. F. de Winter	Delft University of Technology, promotor
Dr. D. Dodou	Delft University of Technology, promotor

Independent members:

Prof. dr. ir. R. Happee	Delft University of Technology
Prof. dr. J. Dankelman	Delft University of Technology
Prof. dr. D. de Waard	University of Groningen
Dr. ir. N. van der Laan	Tilburg University
Dr. ir. M. M. van Paassen	Delft University of Technology
Prof. dr. ir. M. Wisse	Delft University of Technology, reserve member

The majority of research in this dissertation was funded by grant 016.Vidi.178.047 ("How should automated vehicles communicate with other road users?"), provided by the Netherlands Organization for Scientific Research (NWO).

To Amma and Appa

Contents

Summary

A large number of traffic accidents occur worldwide each year, of which a sizable portion involve pedestrians, making them a vulnerable group on the road. Many of these accidents occur due to visual distraction, meaning drivers and pedestrians fail to look at the persons, objects, or locations they should be looking at.

In addition to this tendency for distraction, the eyes are a means of exchanging information between road users, via behaviors such as eye contact. However, the role of eye contact in traffic in connection with traffic safety and the decisions of road users is not yet entirely clear. For example, there is disagreement among experts about the extent to which driver-pedestrian eye contact is important in traffic interactions compared to vehicle motion cues. Further, with the advent of automated vehicles, the role of eye contact in traffic may change or disappear altogether, due to the absence of drivers (or their absence of vehicular control).

One promising way to shed light on this matter might be to use eye-tracking, a technology which can measure the eye movements of road users. Insight into where, when, and how people distribute their attention in traffic might allow the engineering of solutions to mitigate the frequency and severity of accidents. Thus, alongside this technological transformation of traffic, new possibilities arise to boost the safety of vulnerable road users such as pedestrians.

This dissertation aims to investigate the role of eye contact between drivers and pedestrians, as well as its influence on pedestrians' road crossing intentions. Another aim of this dissertation is to assess the accuracy of eye-tracking devices and to objectively detect and operationalize driver-pedestrian eye contact using eye-tracking. Finally, this thesis aims to develop safety systems based on eye-tracking that can automatically analyze and contextualize gaze in traffic and warn vulnerable road users of danger.

This thesis consists of four independently readable and empirical research papers, each presented as their own chapter.

Chapter 2 is an online study conducted among 1835 participants, each of whom observed 13 animated videos twice of an approaching car from the perspective of a pedestrian wanting to cross a road. In some cases, the car yielded to the pedestrian, and in other cases, it did not. A virtual driver in the car made eye contact with the pedestrian at different onset and offset moments, or sometimes avoided eye contact, and participants (pedestrians) could press a key whenever they considered it safe to cross. The results of this study showed that, although the car's kinematics had a dominant effect, eye contact also had an influence on the pedestrians' crossing

intentions. In general, the driver's eye contact made pedestrians feel safer and more likely to cross than an absence of eye contact. Moreover, the onset of the driver's eye contact *alongside* the car's braking led to a substantial increase in willingness to cross. Similarly, the offset of the driver's eye contact alongside the car's acceleration led to a marked decrease in willingness to cross. This study demonstrated the impact of eye contact in traffic and revealed that it is not a binary, static phenomenon, but a time-dependent, dynamic one.

Chapter 3 benchmarks the actual accuracies of eye-tracking glasses in order to lay the foundation for the future objective detection and operationalization of eye contact. While some studies have tested the performance of mobile eye-trackers at different targets at eccentricities, few have measured the effect of dynamic conditions (similar to those experienced by drivers and pedestrians) on their accuracy. This information will be important to put detections of driver-pedestrian eye contact, and gaze on persons and objects (e.g., pedestrians and cars) in the traffic scene, into perspective. In a lab experiment with 36 participants, we assessed two recent mobile eye-trackers, the Tobii Pro Glasses 2 and the Tobii Pro Glasses 3, in scenarios where participants were permitted only eye movements (while seated), only eye and head movements (while seated), or eye, head, and body movements (while walking). Some insights from this study were that eccentricity (i.e., the extent to which the eyes need to rotate) worsened eye-tracking accuracy, but that dynamicity (i.e., the extent to which the user is moving) did not necessarily worsen accuracy. The Tobii 3 was also found to perform better all-round than the Tobii 2. The study also revealed eye-tracking accuracies reported by manufacturers were more favorable than values obtained in practice.

Chapter 4 presents a method to objectively detect and operationalize driver-pedestrian eye contact using two synchronized eye-trackers and computer vision. In a staged experiment involving a driver-pedestrian interaction, 30 participants played the role of a pedestrian standing on a curb or crossing a road in front of a stationary car, while sometimes looking at the driver. Eye contact was detected by tracking the gazes of both driver and pedestrian, and estimating the positions of the two parties with respect to each other via computer vision. The interactions, including the 3D gaze vectors of the persons, were also reconstructed as animations. Driver-pedestrian eye contact was operationalized as both road users simultaneously looking at each other within maximum angular errors of 4° each. This 4° threshold also includes inaccuracies in the eye-trackers themselves. A useful by-product of this study was the technique to automatically analyze gaze data, which bypassed the need for manual annotation in eye-tracking analysis. One future application of the eye contact detection method might be safety systems in automated vehicles that brake when pedestrians or other vulnerable road users have not sought eye contact with the safety driver of the automated vehicle.

Given the above, the question arises whether this knowledge of road user gaze behavior, obtained via eye-tracking, can be used to create safety systems for drivers and pedestrians in the real world. The recent innovations in the field of AI and the availability of vision-language models provide a wealth of opportunities in this regard. **Chapter 5** presents four related concepts of road user safety systems that use combinations of Tobii Pro Glasses 2 eye-tracking, YOLOv8 object detection, and image analysis using GPT-4V, a vision-language model, for the purposes of real-time and automatic analysis of gaze behavior and the estimation of risk in traffic. Such a system would be able to analyze the context of traffic scenes, where the user is (not) looking, and be able to warn the user (e.g., pedestrian, cyclist, or driver) about risky situations. The developed concepts were tested in traffic scenarios of a pedestrian walking in a parking lot, a car driving through urban streets, a distracted pedestrian crossing a street while using a mobile phone, and an additional scenario of a cyclist navigating an urban environment. The results were promising; automatic gaze analysis was achieved in near real-time using YOLOv8, GPT-4V risk ratings of traffic scenes correlated strongly with human risk ratings, and a combination of YOLOv8 and GPT-4V showed potential for analyzing traffic scenes. This study offered real-world evaluations of prototype safety systems for road users, as a first step towards artificially intelligent wearable devices for use in traffic and beyond.

This dissertation **concludes** with a recap of the main results of the four papers and a general discussion of their findings. It argues that while eye contact is neither as powerful a cue as kinematics nor essential for crossing, it is still a “should-have” in driver-pedestrian interactions as it can increase perceived safety and willingness to cross. It therefore concludes that certain types of external Human Machine Interfaces (eHMIs) – substitutes for the missing eye contact between pedestrians and automated vehicles – would be beneficial to maintain existing levels of comfort in interactions. This thesis also advocates the use of (mobile) eye-tracking to detect phenomena such as eye contact in traffic, but cautions against taking accuracy specifications at face value. That said, it operationalizes driver-pedestrian eye contact in a typical interaction as mutual gaze within 4° of each person’s eyes, including any eye-tracker inaccuracies. The dissertation also draws attention to the potential to combine mobile eye-tracking with computer vision and generative artificial intelligence to create context-aware safety systems for road users. It showcases the real-time, human-like capabilities of such systems via four concepts but tempers the former with limitations in terms of analysis quality and operating costs. Finally, this thesis also speculates about future applications of mobile eye-tracking and AI in the traffic, manufacturing, medical, education, and other domains, and recommends topics for further research into eye contact and eye-tracking.

The epilogue briefly covers a related study (in which I was a co-author) that suggests that eye contact in traffic may be an even more intricate phenomenon than previously discussed, especially in traffic cultures where formal rules, e.g., traffic lights and right of way, are not strictly adhered to. An online survey of 3857 drivers found that the reasons for making and avoiding eye contact could be sorted into 15 distinct categories. For example, some drivers deliberately avoid eye contact or pretend not to have seen other nearby drivers in order to manipulate the situation to their advantage and gain right of way. Cultural differences in (the reasons behind) driver-driver eye contact were also found. These final observations provide food for thought regarding the level of social intelligence that future automated vehicles will need to possess in order to navigate complex, mixed traffic effectively, consisting of drivers, pedestrians, and other automated vehicles.

Samenvatting

Wereldwijd gebeuren er elk jaar veel verkeersongevallen, waarvan een aanzienlijk deel voetgangers betreft, een kwetsbare groep op de weg. Veel van deze ongevallen gebeuren door visuele afleiding, wat betekent dat bestuurders en voetgangers niet naar de personen, objecten, of locaties kijken waar ze naar zouden moeten kijken.

Behalve de neiging tot afleiding zijn de ogen ook een middel om informatie uit te wisselen tussen weggebruikers, via gedragingen zoals oogcontact. De rol van oogcontact in het verkeer in relatie tot verkeersveiligheid en de beslissingen die weggebruikers maken is echter nog niet helemaal duidelijk. Er is bijvoorbeeld onenigheid onder experts over de mate waarin oogcontact tussen bestuurder en voetganger belangrijk is in verkeers-interacties in vergelijking met signalen van voertuigbewegingen. Verder kan de rol van oogcontact in het verkeer veranderen of helemaal verdwijnen met de komst van geautomatiseerde voertuigen, vanwege de afwezigheid van bestuurders (of hun afwezigheid van voertuigcontrole).

Een veelbelovende manier om licht op deze kwestie te werpen, zou het gebruik van eye-tracking kunnen zijn, een technologie die de oogbewegingen van weggebruikers kan meten. Inzicht in waar, wanneer en hoe mensen hun aandacht in het verkeer verdelen, kan het mogelijk maken om oplossingen te bedenken om de frequentie en ernst van ongelukken te verminderen. Zo ontstaan er naast deze technologische transformatie van het verkeer nieuwe mogelijkheden om de veiligheid van kwetsbare weggebruikers, zoals voetgangers, te vergroten.

Dit proefschrift heeft als doel de rol van oogcontact tussen bestuurders en voetgangers te onderzoeken, alsmede de invloed ervan op de intenties van voetgangers om de weg over te steken. Een ander doel van dit proefschrift is om de nauwkeurigheid van eye-tracking-apparaten te beoordelen en om oogcontact tussen bestuurder en voetganger objectief te detecteren en operationeel te maken met behulp van eye-tracking. Tot slot heeft dit proefschrift als doel om veiligheidssystemen te ontwikkelen op basis van eye-tracking, welke automatisch iemands blik in het verkeer kunnen analyseren en contextualiseren en kwetsbare weggebruikers kunnen waarschuwen voor gevaar.

Dit proefschrift bestaat uit vier onafhankelijk leesbare en empirische onderzoeksartikelen, elk gepresenteerd als een eigen hoofdstuk.

Hoofdstuk 2 is een online studie uitgevoerd onder 1835 deelnemers, die elk twee keer 13 geanimeerde video's van een naderende auto bekeken vanuit het perspectief van een voetganger die een weg wilde oversteken. In sommige gevallen gaf de auto voorrang aan de voetganger, en in andere gevallen niet. Een virtuele bestuurder in de auto maakte oogcontact met de voetganger op verschillende begin-

en eindmomenten, of vermeed soms oogcontact, en deelnemers (voetgangers) konden op een toets drukken wanneer ze het veilig vonden om over te steken. De resultaten van deze studie toonden aan dat, hoewel de kinematica van de auto een dominant effect had, oogcontact ook invloed had op de oversteekintenties van de voetgangers. Over het algemeen zorgde het oogcontact van de bestuurder ervoor dat voetgangers zich veiliger voelden en eerder geneigd waren om over te steken dan het ontbreken van oogcontact. Bovendien leidde de aanvang van het oogcontact van de bestuurder *tezamen* met het remmen van de auto tot een aanzienlijke toename van de bereidheid om over te steken. Evenzo leidde het einde van het oogcontact van de bestuurder tezamen met de acceleratie van de auto tot een duidelijke afname van de bereidheid om over te steken. Deze studie toonde de impact van oogcontact in het verkeer aan en onthulde dat het geen binair, statisch fenomeen is, maar een tijdsafhankelijk, dynamisch fenomeen.

Hoofdstuk 3 vergelijkt de werkelijke nauwkeurigheid van eye-tracking brillen om de basis te leggen voor de toekomstige objectieve detectie en operationalisering van oogcontact. Hoewel sommige onderzoeken de prestaties van mobiele eye-trackers op verschillende afstanden en met doelen op excentriciteiten hebben getest, hebben er maar weinig het effect van dynamische omstandigheden (vergelijkbaar met die ervaren door bestuurders en voetgangers) op hun nauwkeurigheid gemeten. Deze informatie zal belangrijk zijn om detecties van oogcontact tussen bestuurder en voetganger en blik op personen en objecten (bv. voetgangers en auto's) in het verkeer in perspectief te plaatsen. In een lab-experiment met 36 deelnemers hebben we twee recente mobiele eye-trackers, de Tobii Pro Glasses 2 en de Tobii Pro Glasses 3, beoordeeld in scenario's waarin deelnemers alleen oogbewegingen (zittend), alleen oog- en hoofdbewegingen (zittend), of oog-, hoofd- en lichaamsbewegingen (lopen) mochten maken. Enkele inzichten uit deze studie waren dat excentriciteit (d.w.z. de mate waarin de ogen moeten roteren) de nauwkeurigheid van eye-tracking verslechterde, maar dat dynamiciteit (d.w.z. de mate waarin de gebruiker beweegt) niet noodzakelijkerwijs de nauwkeurigheid verslechterde. De Tobii 3 bleek ook allround beter te presteren dan de Tobii 2. De studie onthulde ook dat de door fabrikanten gerapporteerde nauwkeurigheid van eye-tracking gunstiger was dan de waarden die in de praktijk werden verkregen.

Hoofdstuk 4 presenteert een methode om oogcontact tussen bestuurder en voetganger objectief te detecteren en te operationaliseren met behulp van twee gesynchroniseerde eye-trackers en computer vision. In een 'gestaged' experiment met een interactie tussen bestuurder en voetganger speelden 30 deelnemers de rol van een voetganger die op een stoeprand stond of een weg overstak voor een stilstaande auto, terwijl ze soms naar de bestuurder keken. Oogcontact werd gedetecteerd door de blikken van zowel bestuurder als voetganger te volgen en de posities van de twee partijen ten opzichte van elkaar te schatten via computer vision. De interacties, inclusief de 3D-blikvectoren van de personen, werden ook gereconstrueerd als animaties. Oogcontact tussen bestuurder en voetganger werd

geoperationaliseerd als beide weggebruikers die tegelijkertijd naar elkaar keken binnen maximale hoekafwijkingen van elk 4°. Deze drempel van 4° omdat ook onnauwkeurigheden in de eye-trackers zelf omvatten. Een nuttig bijproduct van deze studie was een techniek om automatisch blikgegevens te analyseren, waardoor de noodzaak van handmatige annotatie bij eye-tracking-analyse werd omzeild. Een toekomstige toepassing van de oogcontactdetectiemethode zou veiligheidssystemen in geautomatiseerde voertuigen kunnen zijn, die remmen wanneer voetgangers of andere kwetsbare weggebruikers geen oogcontact hebben gezocht met de veiligheidsbestuurder van het geautomatiseerde voertuig.

Gegeven het bovenstaande rijst de vraag of deze kennis van het kijkgedrag van weggebruikers, verkregen via eye-tracking, kan worden gebruikt om veiligheidssystemen voor bestuurders en voetgangers in de echte wereld te creëren. De recente innovaties op het gebied van AI en de beschikbaarheid van vision-language-modellen bieden in dit opzicht een schat aan mogelijkheden. **Hoofdstuk 5** presenteert vier gerelateerde concepten van verkeersveiligheidssystemen die combinaties van Tobii Pro Glasses 2 eye-tracking, YOLOv8-objectdetectie en beeldanalyse door GPT-4V, een vision-language model, gebruiken voor realtime en automatische analyse van kijkgedrag en de inschatting van risico's in het verkeer. Een dergelijk systeem zou de context van verkeersscenario's kunnen analyseren, waar de gebruiker (niet) naar kijkt, en de gebruiker (bijvoorbeeld voetganger, fietser of bestuurder) kunnen waarschuwen voor risicovolle situaties. De ontwikkelde concepten werden getest in verkeersscenario's van een voetganger die op een parkeerplaats loopt, een auto die door de stad rijdt, een afgeleide voetganger die een straat oversteekt terwijl hij een mobiele telefoon gebruikt, en een extra scenario van een fietser die door een stedelijke omgeving navigeert. De resultaten waren veelbelovend; automatische kijkgedrag-analyse werd in bijna realtime bereikt met behulp van YOLOv8, GPT-4V-risicobeoordelingen van verkeersscenario's correleerden sterk met menselijke risicobeoordelingen, en een combinatie van YOLOv8 en GPT-4V toonde potentieel voor het analyseren van verkeersscenario's. Deze studie bood real-world evaluaties van prototype veiligheidssystemen voor weggebruikers, als een eerste stap naar kunstmatige intelligente draagbare apparaten voor gebruik in het verkeer en daarbuiten.

Dit proefschrift wordt **afgesloten** met een samenvatting van de belangrijkste resultaten van de vier artikelen en een algemene bespreking van hun bevindingen. Het betoogt dat hoewel oogcontact niet zo'n krachtige cue is als kinematica en ook niet essentieel is voor het oversteken, het nog steeds een 'must-have' is in interacties tussen bestuurder en voetganger, omdat het de waargenomen veiligheid en bereidheid om over te steken kan vergroten. Het proefschrift concludeert daarom dat bepaalde typen external Human Machine Interfaces (eHMIs) – vervangers voor het ontbrekende oogcontact tussen voetgangers en geautomatiseerde voertuigen – gunstig zouden kunnen zijn om de bestaande niveaus van comfort in interacties te behouden. Dit proefschrift pleit ook voor het gebruik van (mobiele) eye-tracking om

verschijnselen zoals oogcontact in het verkeer te detecteren, maar waarschuwt tegen het voor lief nemen van nauwkeurigheids-specificaties. Dat gezegd hebbende, operationaliseert het oogcontact tussen bestuurder en voetganger in een typische interactie als wederzijdse blik binnen 4° van de ogen van elke persoon, inclusief onnauwkeurigheden in de eye-tracker. Het proefschrift vestigt ook de aandacht op het potentieel om mobiele eye-tracking te combineren met computer vision en generatieve kunstmatige intelligentie om context-bewuste veiligheidssystemen voor weggebruikers te creëren. Het toont de real-time, mensachtige mogelijkheden van dergelijke systemen via vier concepten, maar tempert de eerste met beperkingen in termen van analyse-kwaliteit en operationele kosten. Tot slot speculeert dit proefschrift over toekomstige toepassingen van mobiele eye-tracking en AI in het verkeer, fabrieken, de medische sector, het onderwijs en andere domeinen, en beveelt onderwerpen aan voor verder onderzoek naar oogcontact en eye-tracking.

De epiloog behandelt kort een verwant onderzoek (waarvan ik medeauteur was) dat suggereert dat oogcontact in het verkeer een nog ingewikkelder fenomeen kan zijn dan eerder besproken, vooral in culturen waar formele regels, zoals verkeerslichten en voorrang, niet strikt worden nageleefd. Een online enquête onder 3857 bestuurders liet zien dat de redenen om oogcontact te maken en te vermijden in 15 verschillende categorieën konden worden ingedeeld. Sommige bestuurders vermijden bijvoorbeeld opzettelijk oogcontact of doen alsof ze andere bestuurders in de buurt niet hebben gezien om de situatie in hun voordeel te manipuleren en voorrang te krijgen. Er werden ook culturele verschillen in (de redenen achter) oogcontact tussen bestuurders gevonden. Deze laatste observaties bieden stof tot nadenken over het niveau van sociale intelligentie dat toekomstige geautomatiseerde voertuigen moeten bezitten om effectief door complex, gemengd verkeer te navigeren, bestaande uit bestuurders, voetgangers, en andere geautomatiseerde voertuigen.

Chapter 1

Introduction

1.1. Motivation

The primary motivation for this dissertation is to help mitigate the high number of lives lost in traffic each year. This figure was 1.19 million road users in 2023, among whom 23% were pedestrians, numbers that have remained largely unchanged in the last two decades. This is not to mention the 20 to 50 million more people who suffer non-fatal injuries in traffic yearly, many of whom end up with a disability (World Health Organization, 2023).

In the USA, 80% of yearly pedestrian fatalities happen in urban areas (National Highway Traffic Safety Administration, 2023), with approximately three-quarters of those killed in locations other than intersections and without sidewalks (Governors Highway Safety Association, 2023). In over 85% of all cases, pedestrian fatalities are due to frontal impacts by vehicles. Meanwhile, in the Netherlands, in 2022, pedestrian deaths accounted for 8% of all traffic-related fatalities (SWOV, 2023), and roughly three-quarters of those occurred in urban areas (European Road Safety Observatory, 2023).

A pattern can be seen here: pedestrians are most often killed in frontal collisions with vehicles in urban settings, a problem that is exacerbated when there is a lack of road safety infrastructure (e.g., traffic lights, zebra crossings, sidewalks) and/or when traffic rules such as right of way are ambiguous. Many explanations for these casualties have been offered by the aforementioned reports, including inattention and distraction, for example due to mobile phones, in both drivers and pedestrians. That being said, a major consequence of such behaviors is an inability or a failure to communicate properly with the other road user.

Several studies note that communication in traffic is a critical aspect of road safety (Ackermann et al., 2019; Kong et al., 2021; Zandi et al., 2020). Since communication almost always happens via either implicit or non-verbal cues (Lee et al., 2021; Rasouli et al., 2017), these behaviors are worth investigating. Markkula et al. (2020) define implicit communication as “a road user behavior which affects own movement or perception, but which can at the same time be interpreted as signaling something to or requesting something from another road user” (p. 741). Examples of implicit communication include pedestrian and vehicle orientation and motion, and gaze direction. While there is some overlap between these and non-verbal cues by definition, the latter typically refer to explicit communication or communicative actions directed specifically at other road users. Markkula et al. (2020) define explicit communication as “a road user behavior which *does not* affect own movement or perception, but which can be interpreted as signaling something to or requesting something from another road user” (p. 742; Italics added). Examples of this include eye contact, nodding, and hand gestures (Färber, 2016).

Coming to the subject of gaze, a person’s gaze behavior may be considered the pattern of their gaze directions and durations in a given environment. The gaze

behavior of road users can offer a window into their minds, because it has been observed that there is a strong correlation between where a person is looking and what they are thinking about (Just & Carpenter, 1980). While, in isolation, gaze behavior on the road is an implicit cue (e.g., a pedestrian checking traffic lights before crossing a street), it can become an explicit cue when directed at other road users, e.g., a pedestrian looking at a driver before crossing a street.

Eye contact is a special instance of gaze behavior and is particularly interesting due to its simultaneous potential for clarity and ambiguity as a cue. Past research has generally shown that eye contact, much like other types of road communication, can be beneficial for road safety, or at least perceived road safety (Habibovic et al., 2018; Malmsten Lundgren et al., 2017; Ren et al., 2016; Sucha et al., 2017). This claim is not without its detractors, however, and a few studies dispute its usefulness or even its occurrence (AlAdawy et al., 2019; Dey & Terken, 2017; Moore et al., 2019).

Eye contact is also a challenging behavior to decipher. In an instant, it is able to convey the intentions and determination of a road user to another, and yet can also signal uncertainty and indecisiveness. Moreover, the psychological processes that underpin eye contact in traffic are complex and not yet fully understood. For instance, it is not entirely clear what goes through the minds of a driver and a pedestrian when they make eye contact as they simultaneously approach a junction, which leads to a crossing conflict and requires negotiation of right of way.

An extra dimension of complexity is that eye contact often occurs in combination with other cues like pedestrian and vehicle motion, and its duration and time of onset vary widely from one road interaction to the next. It also is a two-way phenomenon, dependent on both driver and pedestrian to engage one another at the same time. All these factors make eye contact in traffic difficult to isolate and operationalize.

Finally, with the advent of automated vehicles (AVs) and the entire or partial removal of drivers from the vehicle control loop, effectively becoming passengers or stand-by operators, more often inattentive or altogether absent, concerns about reductions in pedestrian safety arise (Färber, 2016; Habibovic et al., 2018; Malmsten Lundgren et al., 2017). This is primarily due to the ineffectiveness of eye contact (and other forms of non-verbal communication) in the traditional sense, since the AV's "driver" (occupant) is either incapable of acknowledging and reciprocating the pedestrian's non-verbal cues or is capable but this has no effect on the vehicle's actions. Therefore, in preparation for this new era of traffic interactions, it is prudent to research and objectively describe road user behaviors like eye contact, in the interest of retaining or improving any safety benefits they bring, especially in a future when the behaviors themselves may no longer be able to occur.

1.2. Research gap

Traffic studies often overlook the aforementioned complexities in eye contact and tend to reduce it to a one-sided, standalone, instantaneous, or binary action (Malmsten Lundgren et al., 2017; Ren et al., 2016; Sucha et al., 2017). In other words, they often record it from only one perspective, i.e., either the driver's or the pedestrian's, view it as independent from other cues, reduce it to one defining moment, or treat it as a dichotomy, i.e., simply either present or absent in a road interaction.

These drawbacks raise the question of what methods and tools to use to best capture eye contact. Among some of the more recent studies, Lanzer et al. (2021) noted, based on manual annotation of naturalistic vehicle-pedestrian interaction videos recorded from the vehicle's perspective, that about 63% of pedestrians gazed at least once in the direction of the approaching vehicle. Although this did not necessarily imply eye contact since the pedestrians' exact gaze locations were unknown, it did highlight the difficulty in differentiating eye contact from glances at the vehicle or the environment when reviewing video footage.

Belkada et al. (2024) introduced body, head, and eye features in a deep learning model to detect pedestrians' eye contact attempts from images of driving scenes borrowed from existing autonomous driving datasets. They found that the model achieved the best performance when using features around the eyes only. The authors also provided annotations of 57 thousand instances of pedestrians' looking behavior from three autonomous driving datasets. While this likely aids automatic detection in future studies, the dataset suffers from the same uncertainty pitfall of whether the pedestrians were really making eye contact.

The same problem even affected studies that employed simultaneous direct observation of drivers and pedestrians interacting in real traffic, ultimately leading the researchers to remove eye contact from their observation protocol, as it was too difficult to determine with confidence (Lee et al., 2021). Earlier works using camera recordings and direct on-site observations of driver-pedestrian interactions fared similarly, encountering the same uncertainty problem (Sucha et al., 2017), with some of the papers advising the operationalization of eye contact as (at least) an arbitrary number of seconds of gaze in the direction of a road user (Rasouli et al., 2017). All these endeavors point to the larger problem of accurately determining the target object and duration of a road user's gaze, a phenomenon of which eye contact is but one instance.

One promising technique that does not face the above uncertainty pitfall, is resistant to timing variations, and also does not rely on subjective self-reports of eye contact by drivers and pedestrians, is eye-tracking. This technology has been used in automotive research for decades due to its ability to accurately localize a person's gaze (Land & Lee, 1994; Mourant & Rockwell, 1972; Sodhi et al., 2002). Some of

these devices worked in real-time, were head-mounted, and in theory, portable, which made naturalistic studies of drivers in environments such as vehicle cockpits possible. However, the complexity of the eye-tracker setups and the need for cables and auxiliary equipment were drawbacks that hindered naturalistic studies of pedestrians on the road.

More recently, the development of truly portable eye-trackers (often in the form of wearable glasses) that are mobile, lightweight, wireless, and require minimum setup has revolutionized traffic research by their ability to be used in outdoor environments, dynamic conditions, and tight spaces (Babić et al., 2020; Fotios et al., 2015; Gruden et al., 2021; Zito et al., 2015). With this have come investigations into driver-pedestrian eye contact, within the larger realm of studies on road user gaze behavior. For example, Nathanael et al. (2019) tracked the eyes of drivers on urban roads using a Tobii Pro Glasses 2 eye-tracker, and found that pedestrians gazed towards the vehicle in 65% of all interactions, including 11% in which eye contact occurred, i.e., the drivers gazed back at the pedestrians' faces. The authors noted that in almost all of the latter cases, the crossing conflict was resolved solely on the basis of eye contact, without the need for additional cues. Similarly, De Winter et al. (2021) used the Tobii Pro Glasses 2 and observed that pedestrians navigating a parking lot sought eye contact with drivers of parked and moving cars for a median of 4.5% of the total experiment recording time, and that in 25% of cases where pedestrians looked in the direction of an approaching vehicle, they looked at the driver.

In spite of mobile eye-tracking being a potentially effective tool to study the gaze behaviors (including eye contact) of road users, the technology suffers from a few problems:

1. Compared to remote eye-trackers, wearable eye-tracking glasses encounter a number of challenges which might undermine their accuracy:
   a. Being used in conditions that are dynamic, which risks that the glasses slip from the position on the user's face at which they were calibrated (Niehorster et al., 2020), have varying lighting conditions and disturbances from sunlight (e.g., Tatler et al., 2019), and involve different/larger target distances than the distance used for calibration (MacInnes et al., 2018).
   b. Less rigorous calibration procedures for the sake of user-friendliness (e.g., one-point calibration in the Tobii Pro Glasses 2 and 3 versus nine-point calibration in SR Research's EyeLink 1000 Plus);
   c. Lower sampling frequencies than those in remote eye-trackers (e.g., 50 or 100 Hz in the Tobii Pro Glasses models versus 2000 Hz in the EyeLink);
   d. Involuntary movement during calibration because of being head-mounted.
2. Just one eye-tracker is insufficient to fully capture a phenomenon that involves the simultaneous gazes of multiple persons (e.g., joint attention or eye contact in traffic).

3. There is currently no easy way to automate both the analysis and contextualization of mobile eye-tracking data with respect to video footage of a user's point of view, especially in real-time (Barz & Sonntag, 2021; Fong et al., 2016). In other words, it is hard to determine automatically, i.e., without resorting partly or wholly to manual, post-hoc review, what object/person/environment features a user is looking at and what such gaze might mean in that context.

1.3. Research goal

Thus, in view of the research gaps, the goals of this PhD thesis are fourfold:

1. Investigate the effect of eye contact in traffic on road user behavior, while considering eye contact as a complex and variable cue.
2. Benchmark the accuracy of mobile eye-trackers under static and dynamic conditions.
3. Operationalize driver-pedestrian eye contact fully using two eye-trackers.
4. Develop a method to automatically analyze and contextualize mobile eye-tracking data in traffic scenarios, preferably in real-time.

To elaborate, this thesis aims to provide a deeper understanding of the role and importance (if any) of eye contact between road users in traffic interactions, preferably with an emphasis on pedestrians, the most vulnerable road users (SWOV, 2023; World Health Organization, 2023). Next, it tries to quantify how reliable the gaze measurements of mobile eye-trackers are under various conditions, as a prelude to using the devices in naturalistic traffic settings. That done, it strives to push the boundaries of using eye-tracking technology to operationalize driver-pedestrian eye contact, in order to better understand this communication strategy. Finally, it seeks to take the first steps towards developing systems that can automatically and in real-time, make sense of mobile eye-tracking data. The ultimate goal is to lay the groundwork for the future development of eye-tracking-based assistance and safety systems that can assess risk and provide gaze-contingent and context-specific feedback to a user (e.g., warnings, information bulletins), in a variety of environments.

Based on the above goals and the automotive scope of this thesis, the main research questions may be formulated, respectively, as follows:

1. What is the effect of eye contact, its timing, and its duration on the behavior of pedestrians in a traffic interaction?
2. How accurate are mobile eye-trackers under various levels of dynamicity?
3. How can eye contact between a driver and a pedestrian be detected and operationalized objectively?
4. How can eye-tracking data of road users be automatically analyzed and contextualized for the purpose of traffic risk assessment and the development of safety systems?

A wide variety of methods are used to answer these research questions, ranging from questionnaires and crowdsourced, online experiments, to indoor re-creations of traffic scenarios and outdoor, naturalistic trials. This broad spectrum of techniques aims to lend validity to the findings in both real-world traffic applications and future laboratory-based research.

1.4. Thesis outline

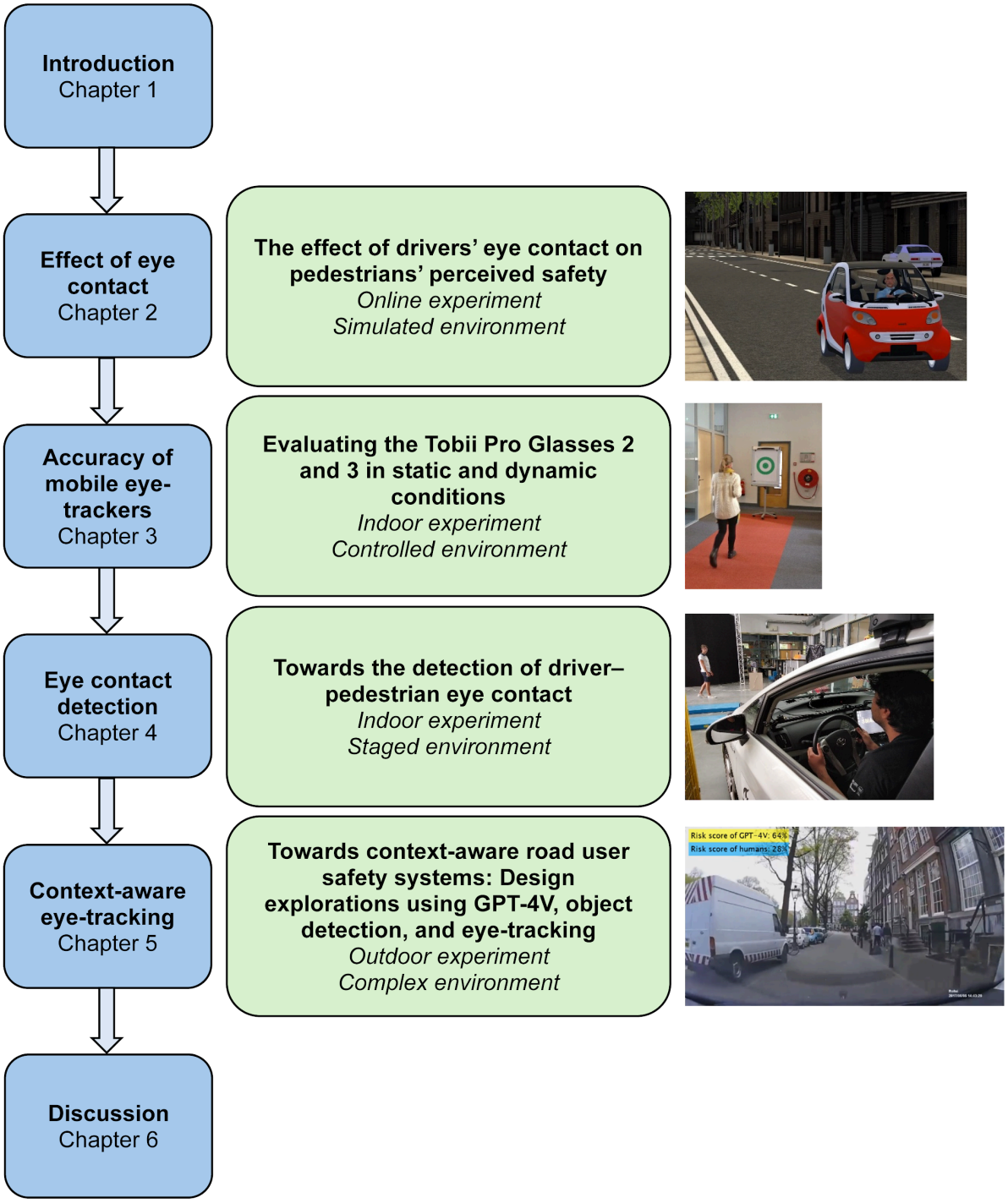

Figure 1.1. Outline of the thesis chapters and their corresponding research papers.

This thesis is divided into six chapters, as shown in Figure 1.1. First, this introduction chapter provides a brief overview of the state of the art in eye contact and eye-tracking research in the traffic context. It also identifies research gaps in the literature, outlines the research goals of this thesis, and poses research questions.

Chapter 2 contains a journal paper that explores the role of drivers' eye contact in pedestrians' feeling of safety to cross a road in front of an approaching car. The experiment was conducted online in a crowdsourced manner with a large sample size, using videos of a driver in an oncoming car as seen from a pedestrian's perspective, generated from a simulated environment. The study is innovative in its treatment of eye contact as a nuanced, dynamic variable, i.e., not simply "present" or "absent" in a given interaction, but instead varying in its timing and duration, with different possible start and end points during the crossing conflict. The study is also innovative in its isolation of eye contact from other confounding driver states, such as distraction, drowsiness, or absence. Only the car's yielding behavior was retained as an additional variable, allowing insight into the relative strength of drivers' eye contact and vehicle motion as cues for pedestrians. Questionnaires were administered before and after the experiment to ascertain the validity of the findings against the opinions of pedestrians on the importance and usefulness of drivers' eye contact for pedestrian safety. In this manner, Chapter 2 answers the first research question of this thesis on the role of eye contact in traffic. This chapter also reiterates safety concerns surrounding the potential disappearance of non-verbal cues like eye contact in future pedestrian interactions with automated vehicles.

The results of Chapter 2 may find application in the development of AVs and in the design of traffic safety infrastructure, particularly about when and how these entities should communicate with pedestrians and other vulnerable road users (VRUs), informed by a better understanding of eye contact in traffic.

Chapter 3 presents a journal paper that evaluates the eye-tracking accuracy of two popular mobile eye-trackers, the Tobii Pro Glasses 2 and the Tobii Pro Glasses 3 (occasionally referred to simply as Tobii 2 and Tobii 3 in this thesis), under multiple levels of dynamic conditions. This paper forms an intermediate but necessary step by benchmarking the capabilities of modern eye-tracking glasses in situations with different levels of dynamicity, so that eye-tracking data may be interpreted correctly. The study acts as a precursor to future efforts to detect driver-pedestrian eye contact using mobile eye-tracking. While manufacturer-quoted accuracy specifications do exist, past research has found discrepancies between these reported values and the researchers' own observed values, with the latter accuracies often being poorer (Ehinger et al., 2019; Holmqvist, 2017; Stuart et al., 2016). Chapter 3 is innovative in its assessment of accuracy as a function of dynamicity, as opposed to target eccentricity, target distance, or ambient illumination level, which are more common approaches in eye-tracker testing (MacInnes et al., 2018; Tatler et al., 2019). The experiment was performed in a laboratory setting with a moderate sample size of

participants who were, at times, allowed to (1) move only their eyes, (2) move only their eyes and their head, or (3) move their eyes, head, and body, while gazing at designated targets. In this way, Chapter 3 addresses the second research question of this thesis on the accuracy of the chosen tools for detecting eye contact. The tests in this chapter may also, by extension, provide insight into the reliability of mobile eye-tracking when dealing with head movements and eye movements of drivers and pedestrians in the real world.

The applications of Chapter 3 lie in future advancements in eye-tracking technology, where the current findings may help develop more accurate and precise eye-trackers, especially in dynamic conditions or when gazing at eccentric targets. Such improvements will likely benefit research in various fields, including traffic safety, psychology, marketing, sports, video gaming, art, and human-computer interaction.

The journal paper that comprises Chapter 4 proposes a new method to objectively detect and define driver-pedestrian eye contact using two synchronized eye-trackers, two cameras, and image recognition. It fills a major gap in the literature on eye contact measurement techniques by eliminating subjective, indirect, “third-person”, or one-sided assessments of eye contact between drivers and pedestrians, e.g., direct observations by researchers, reviews of traffic camera footage, self-reports by road users, estimations from road user head orientations, and eye-tracking of only one of the two parties. The study makes the technological leap from merely being able to infer eye contact between two road users to definitively detecting it, and validates this via a staged, indoor experiment with a moderate sample size that recreates a driver-pedestrian interaction involving eye contact. The paper operationalizes eye contact in a typical driver-pedestrian interaction as both persons looking at each other’s eyes simultaneously, within an error of 4° in gaze directions. In this manner, Chapter 4 answers the third research question of this thesis on how to detect and operationalize driver-pedestrian eye contact. The objective and precise definition may aid in the development of safety solutions to compensate for the lack of eye contact in pedestrian interactions with automated vehicles in the future.

The findings of Chapter 4 are applicable to future efforts in modeling driver-pedestrian interactions and non-verbal communication in traffic. The method proposed could form the basis for safety systems incorporating eye-tracking and eye contact and operating from the pedestrian’s, driver’s, or vehicle’s perspective. Examples might include wearable devices and sensor modules in automated vehicles for detecting, interpreting, and if necessary, responding to eye contact (or more generally, non-verbal communication) between road users. Additionally, similar applications may be found in other areas where accurate detection of eye contact is useful, e.g., social interaction, education.

Chapter 5 contains a paper that explores the possibility to use eye-tracking to create a wearable, context-aware safety system for road users, by combining the former with object detection and generative AI. The study presents a demonstration of such a system that uses combinations of Tobii Pro Glasses 2 eye-tracking, the object-detection method YOLOv8, and the vision-language model GPT-4V. Test runs were performed in four distinct scenarios, ranging from an indoor scene observation and a pedestrian navigating a parking lot to a driver in an urban environment and a distracted pedestrian crossing a road while using a mobile phone. One novelty of this endeavor is the automatic context analysis of eye-tracking video and gaze data, eliminating the need for labor-intensive post-hoc annotation, environmental markers, and manual review of footage. Another is the provision of context-specific feedback to the user about instantaneous risk, in the form of a rating (0–100), and brief text about the user's relation to the immediate environment and the biggest safety risks in the moment, both of which aim to increase the user's situational awareness. Since this system is intended as a proof-of-concept, tests were restricted to a small number of trials in naturalistic environments. In this way, Chapter 5 answers the fourth and last research question of this thesis on automatically analyzing eye-tracking data and estimating traffic risk in order to develop safety systems. On a broader scale, Chapter 5 strengthens the applicability of eye-tracking technology to real-world applications in the field of traffic safety.

In a similar vein as the previous chapter, the applications of Chapter 5 lie in the future development of real-time, wearable, and artificially intelligent assistance systems that can provide safety warnings, information bulletins, and context-appropriate feedback in a variety of settings, e.g., traffic (walking, cycling, driving), industrial work, social interaction, shopping, household chores, and human-computer interaction. The integration of eye-tracking with object detection and AI analysis can help raise situational awareness and aid in tasks requiring rapid contextualization of user focus in scenes.

Finally, Chapter 6 recaps the main findings of all the preceding chapters of this thesis, and discusses the conclusions drawn from them. It also provides an outlook on the future of eye contact and eye-tracking in the traffic context and beyond, and offers recommendations for follow-up research. In the epilogue, Chapter 6 includes a brief discussion of a related, fifth paper in which I was a co-author.

References

Ackermann, C., Beggiato, M., Bluhm, L.-F., Löw, A., & Krems, J. F. (2019). Deceleration parameters and their applicability as informal communication signal between pedestrians and automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour, 62*, 757–768. https://doi.org/10.1016/j.trf.2019.03.006

AlAdawy, D., Glazer, M., Terwilliger, J., Schmidt, H., Domeyer, J., Mehler, B., Reimer, B., & Fridman, L. (2019). Eye contact between pedestrians and drivers.

Proceedings of the Tenth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Santa Fe, New Mexico, 301–307. https://pubs.lib.uiowa.edu/driving/article/28343/galley/136635/view

Babić, D., Dijanić, H., Jakob, L., Babić, D., & Garcia-Garzon, E. (2020). Driver eye movements in relation to unfamiliar traffic signs: An eye tracking study. *Applied Ergonomics, 89*, Article 103191. https://doi.org/10.1016/j.apergo.2020.103191

Barz, M., & Sonntag, D. (2021). Automatic visual attention detection for mobile eye tracking using pre-trained computer vision models and human gaze. *Sensors, 21*, Article 4143. https://doi.org/10.3390/s21124143

Belkada, Y., Bertoni, L., Caristan, R., Mordan, T., & Alahi, A. (2024). *Do pedestrians pay attention? Eye contact detection in the wild*. SSRN. https://doi.org/10.2139/ssrn.4760697

De Winter, J., Bazilinskyy, P., Wesdorp, D., De Vlam, V., Hopmans, B., Visscher, J., & Dodou, D. (2021). How do pedestrians distribute their visual attention when walking through a parking garage? An eye-tracking study. *Ergonomics, 64*, 793–805. https://doi.org/10.1080/00140139.2020.1862310

Dey, D., & Terken, J. (2017). Pedestrian interaction with vehicles: Roles of explicit and implicit communication. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Oldenburg, Germany, 109–113. https://doi.org/10.1145/3122986.3123009

Ehinger, B. V., Groß, K., Ibs, I., & König, P. (2019). A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. *PeerJ, 7*, Article e7086. https://doi.org/10.7717/peerj.7086

European Road Safety Observatory. (2023). National road safety profile - Netherlands. https://road-safety.transport.ec.europa.eu/system/files/2023-02/erso-country-overview-2023-netherlands_0.pdf

Färber, B. (2016). Communication and communication problems between autonomous vehicles and human drivers. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.) *Autonomous driving* (pp. 122–144). Springer. https://doi.org/10.1007/978-3-662-48847-8_7

Fong, A., Hoffman, D., & Ratwani, R. M. (2016). Making sense of mobile eye-tracking data in the real-world: A human-in-the-loop analysis approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*, 1569–1573. https://doi.org/10.1177/1541931213601362

Fotios, S., Uttley, J., Cheal, C., & Hara, N. (2015). Using eye-tracking to identify pedestrians' critical visual tasks, Part 1. Dual task approach. *Lighting Research & Technology, 47*, 133–148. https://doi.org/10.1177/1477153514522472

Governors Highway Safety Association. (2023). Pedestrian traffic fatalities by state: 2022 preliminary data. https://www.ghsa.org/resources/Pedestrians23

Gruden, C., Ištoka Otković, I., & Šraml, M. (2021). Safety analysis of young pedestrian behavior at signalized intersections: An eye-tracking study. *Sustainability, 13*, Article 4419. https://doi.org/10.3390/su13084419

Habibovic, A., Lundgren, V. M., Andersson, J., Klingegård, M., Lagström, T., Sirkka, A., Fagerlönn, J., Edgren, C., Fredriksson, R., Krupenia, S., Saluäär, D., & Larsson, P. (2018). Communicating intent of automated vehicles to pedestrians. *Frontiers in Psychology, 9*, Article 1336. https://doi.org/10.3389/fpsyg.2018.01336

Holmqvist, K. (2017). *Common predictors of accuracy, precision and data loss in 12 eye-trackers*. ResearchGate. https://doi.org/10.13140/RG.2.2.16805.22246

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Kong, X., Das, S., Zhang, Y., & Xiao, X. (2021). Lessons learned from pedestrian-driver communication and yielding patterns. *Transportation Research Part F: Traffic Psychology and Behaviour, 79*, 35–48. https://doi.org/10.1016/j.trf.2021.03.011

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature, 369*, 742–744. https://doi.org/10.1038/369742a0

Lanzer, M., Gieselmann, M., Mühl, K., & Baumann, M. (2021). Assessing crossing and communication behavior of pedestrians at urban streets. *Transportation Research Part F: Traffic Psychology and Behaviour, 80*, 341–358. https://doi.org/10.1016/j.trf.2021.05.001

Lee, Y. M., Madigan, R., Giles, O., Garach-Morcillo, L., Markkula, G., Fox, C., Camara, F., Rothmueller, M., Vendelbo-Larsen, S. A., Rasmussen, P. H., Dietrich, A., Nathanael, D., Portouli, V., Schieben, A., & Merat, N. (2021). Road users rarely use explicit communication when interacting in today's traffic: Implications for automated vehicles. *Cognition, Technology & Work, 23*, 367–380. https://doi.org/10.1007/s10111-020-00635-y

MacInnes, J. J., Iqbal, S., Pearson, J., & Johnson, E. N. (2018). *Wearable eye-tracking for research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices*. BioRxiv. https://doi.org/10.1101/299925

Malmsten Lundgren, V., Habibovic, A., Andersson, J., Lagström, T., Nilsson, M., Sirkka, A., Fagerlönn, J., Fredriksson, R., Edgren, C., Krupenia, S., & Saluäär, D. (2017). Will there be new communication needs when introducing automated vehicles to the urban context? In N. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation* (pp. 485–497). Springer. https://doi.org/10.1007/978-3-319-41682-3_41

Markkula, G., Madigan, R., Nathanael, D., Portouli, E., Lee, Y. M., Dietrich, A., Billington, J., Schieben, A., & Merat, N. (2020). Defining interactions: A conceptual framework for understanding interactive behaviour in human and automated road traffic. *Theoretical Issues in Ergonomics Science, 21*, 728–752. https://doi.org/10.1080/1463922x.2020.1736686

Moore, D., Currano, R., Strack, G. E., & Sirkin, D. (2019). The case for implicit external human-machine interfaces for autonomous vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive*

Vehicular Applications, Utrecht, the Netherlands, 295–307. https://doi.org/10.1145/3342197.3345320

Mourant, R. R., & Rockwell, T. H. (1972). Strategies of visual search by novice and experienced drivers. *Human Factors, 14*, 325–335. https://doi.org/10.1177/001872087201400405

Nathanael, D., Portouli, E., Papakostopoulos, V., Gkikas, K., & Amditis, A. (2019). Naturalistic observation of interactions between car drivers and pedestrians in high density urban settings. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association* (pp. 389–397). Springer. https://doi.org/10.1007/978-3-319-96074-6_42

National Highway Traffic Safety Administration. (2023). Traffic safety facts: 2021 data. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813458

Niehorster, D. C., Santini, T., Hessels, R. S., Hooge, I. T. C., Kasneci, E., & Nyström, M. (2020). The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods, 52*, 1140–1160. https://doi.org/10.3758/s13428-019-01307-0

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Agreeing to cross: How drivers and pedestrians communicate. *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium*, Los Angeles, CA, 264–269. https://doi.org/10.1109/IVS.2017.7995730

Ren, Z., Jiang, X., & Wang, W. (2016). Analysis of the influence of pedestrians' eye contact on drivers' comfort boundary during the crossing conflict. *Procedia Engineering, 137*, 399–406. https://doi.org/10.1016/j.proeng.2016.01.274

Sodhi, M., Reimer, B., Cohen, J. L., Vastenburg, E., Kaars, R., & Kirschenbaum, S. (2002). On-road driver eye movement tracking using head-mounted devices. *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, New Orleans, Louisiana, 61–68. https://doi.org/10.1145/507072.507086

Stuart, S., Alcock, L., Godfrey, A., Lord, S., Rochester, L., & Galna, B. (2016). Accuracy and re-test reliability of mobile eye-tracking in Parkinson's disease and older adults. *Medical Engineering & Physics, 38*, 308–315. https://doi.org/10.1016/j.medengphy.2015.12.001

Sucha, M., Dostal, D., & Risser, R. (2017). Pedestrian-driver communication and decision strategies at marked crossings. *Accident Analysis & Prevention, 102*, 41–50. https://doi.org/10.1016/j.aap.2017.02.018

SWOV. (2023, September 23). Road deaths in the Netherlands: SWOV fact sheet, September 2023. https://web.archive.org/web/20240103154323/https://swov.nl/en/fact-sheet/road-deaths-netherlands

Tatler, B. W., Hansen, D. W., & Pelz, J. B. (2019). Eye movement recordings in natural settings. Eye movement research: An introduction to its scientific foundations and applications. In C. Klein & U. Ettinger (Eds.), *Eye movement research. Studies in neuroscience, psychology and behavioral economics* (pp. 549–592). Springer. https://doi.org/10.1007/978-3-030-20085-5_13

World Health Organization. (2023, December 13). Global status report on road safety 2023. https://www.who.int/publications/i/item/9789240086517

Zandi, B., Singer, T., Kobbert, J., & Quoc Khanh, T. (2020). International study on the importance of communication between automated vehicles and pedestrians. *Transportation Research Part F: Traffic Psychology and Behaviour, 74*, 52–66. https://doi.org/10.1016/j.trf.2020.08.006

Zito, G. A., Cazzoli, D., Scheffler, L., Jäger, M., Müri, R. M., Mosimann, U. P., Nyffeler, T., Mast, F. W., & Nef, T. (2015). Street crossing behavior in younger and older pedestrians: An eye- and head-tracking study. *BMC Geriatrics, 15*, Article 176. https://doi.org/10.1186/s12877-015-0175-0

Chapter 2

The effect of drivers' eye contact on pedestrians' perceived safety

This chapter has been published as:
Onkhar, V., Bazilinskyy, P., Dodou, D., & De Winter, J. C. F. (2022). The effect of drivers' eye contact on pedestrians' perceived safety. *Transportation Research Part F: Traffic Psychology and Behaviour, 84*, 194–210. https://doi.org/10.1016/j.trf.2021.10.017

Abstract

Many fatal accidents that involve pedestrians occur at road crossings, and are attributed to a breakdown of communication between pedestrians and drivers. Thus, it is important to investigate how forms of communication in traffic, such as eye contact, influence crossing decisions. Thus far, there is little information about the effect of drivers' eye contact on pedestrians' perceived safety to cross the road. Existing studies treat eye contact as immutable, i.e., it is either present or absent in the whole interaction, an approach that overlooks the effect of the timing of eye contact. We present an online crowdsourced study that addresses this research gap. 1835 participants viewed 13 videos of an approaching car twice, in random order, and held a key whenever they felt safe to cross. The videos differed in terms of whether the car yielded or not, whether the car driver made eye contact or not, and the times when the driver made eye contact. Participants also answered questions about their perceived intuitiveness of the driver's eye contact behavior. The results showed that eye contact made people feel considerably safer to cross compared to no eye contact (an increase in keypress percentage from 31% to 50% was observed). In addition, the initiation and termination of eye contact affected perceived safety to cross more strongly than continuous eye contact and a lack of it, respectively. The car's motion, however, was a more dominant factor. Additionally, the driver's eye contact when the car braked was considered intuitive, and when it drove off, counterintuitive. In summary, this study demonstrates for the first time how drivers' eye contact affects pedestrians' perceived safety as a function of time in a dynamic scenario and questions the notion in recent literature that eye contact in road interactions is dispensable. These findings may be of interest in the development of automated vehicles (AVs), where the driver of the AV might not always be paying attention to the environment.

2.1. Introduction

Worldwide, more than 50% of traffic-related deaths are that of vulnerable road users such as pedestrians (World Health Organization, 2020). Most pedestrian deaths occur in urban areas at non-intersection locations (National Highway Traffic Safety Administration, 2020; SWOV, 2020).

A possible cause of these casualties is a breakdown in communication with other road users such as car drivers (European Road Safety Observatory, 2018). Färber (2016) noted that road users communicate via informal means, such as eye contact, in addition to relying on formal traffic rules. Understanding the role of eye contact in traffic is a relevant topic in recent times, with the development of AVs in which the driver may be intermittently attentive. To illustrate, according to Google Scholar, from the 45 papers citing a recent paper on the effect of pedestrian's eye contact on the speed of approaching vehicles (Ren et al., 2016), 30 (67%) are directly related to AV interaction with vulnerable road users, based on the titles of the citing works.

Thus far, a few studies have investigated the effect of drivers' eye contact on pedestrians, as listed below:

- In a Wizard of Oz experiment, Malmsten Lundgren et al. (2017; *N* = 13) found that pedestrians reported greater willingness to cross the road when the driver of a supposed AV made eye contact with them, compared to when the driver was inattentive by reading a newspaper or talking on the phone.
- Yang (2017; *N* = 40) presented participants with pictures of a driver making eye contact, talking on the phone, sleeping, or being hidden from view by blinded windows. This study found that the driver's eye contact made participants feel more certain they were safe to cross compared to when the driver was inattentive or hidden from view.
- In a crowdsourcing study using images of an intersection from a cyclist's perspective, Bazilinskyy et al. (2023) found that a driver's eye contact increased the willingness of cyclists to cross compared to no driver's eye contact. This effect was only found in their second experiment (*N* = 1086), in which observers were asked to rate features of an AV in the image; it was not found in their first experiment (*N* = 1260), in which participants made a quick go/stop decision.
- In a Wizard of Oz study by Rodríguez Palmeiro et al. (2018; *N* = 24), no significant differences were observed between pedestrians' moments of making a crossing decision between attentive-driver and distracted-driver conditions, including a distracted driver reading a newspaper.
- In another Wizard of Oz study by Faas et al. (2021; *N* = 65), pedestrians felt safer to cross in front of a car with a driver making eye contact compared to a driver reading a newspaper or a car with blinded windows. No significant differences were observed, however, in terms of crossing onset times.
- In a study using a head-mounted display, Núñez Velasco et al. (2021; *N* = 20) let pedestrians cross a virtual road in front of an AV with an external Human Machine Interface (eHMI), which featured an attentive driver, a distracted driver, or no driver. The study concluded that "*the most important factor affecting pedestrians' road crossing behavior was the motion cues derived from the vehicle, rather than the presence or state of the driver. This raises the question about the needs, purpose, and added value of eHMIs*" (p. 57).
- In a virtual reality study, Chang et al. (2017; *N* = 15) evaluated an eHMI in the form of artificial eyes and found that pedestrians reached a correct crossing decision faster and reported feeling safer when the eyes made eye contact with them, compared to when they did not. Similar eHMI concepts were proposed by Alvarez et al. (2019, 2020), Jaguar Land Rover (2018), Löcken et al. (2019), Mahadevan et al. (2018), Pennycooke (2012), Verma et al. (2019), and Wang et al. (2021). In Löcken et al. (2019), the virtual eyes concept came out as the most untrustworthy from a total of five eHMI concepts. Additionally, Furuya et al. (2021) tested a virtual human embodiment in AV-pedestrian interaction and found that a 'driver that looks at you' was preferred over 'no driver' and a 'static driver' by 25 out of 26 participants.

From the above, it appears that drivers' eye contact can, in some cases, make pedestrians feel safer to cross as compared to no eye contact. These findings are in line with research in social and evolutionary psychology, suggesting that eye contact has various functional rules, including signaling, recognizing, facilitating joint attention, and encouraging compliance (Argyle & Dean, 1965; Hamlet et al., 1984; Tomasello et al., 2007).

However, most of the studies listed earlier concerned situations where the driver was completely disengaged from the driving task, for example by reading a newspaper. These are unlikely scenarios to encounter on real roads since AVs are not yet at a level of automation that (legally) allows drivers to be this lax at the wheel. Thus, the positive impact of drivers' eye contact on pedestrians' crossing behavior noted by the above studies may be because these studies used a completely disengaged driver as the baseline, instead of simply an attentive driver who does not make eye contact.

More problematically, the above studies used only two simplified conditions: eye contact is either present or absent in the crossing conflict. They also used only a single go/stop decision moment without examining the evolution of such decision-making as the car is approaching. Since eye contact is a phenomenon that spans a finite length of time, and because traffic interactions themselves are typically brief affairs, there is incentive to investigate eye contact in relation to crossing behavior as a function of time. The results of such an approach would provide a more truthful account of the importance of eye contact on the road.

At the same time, it has been argued that implicit communication cues, viz. car speed and distance are probably more dominant cues for pedestrians to understand the intention of an approaching car (Clamann et al., 2017; Dey & Terken, 2017; Lee et al., 2021; Núñez Velasco et al., 2021). In an online survey study, AlAdawy et al. (2019) reported that pedestrians are usually unable to see the driver through the windshield because of sunshine, shadows, glare, or darkness. In the same vein, it has been noted that drivers are less compelled than pedestrians to make eye contact (Sucha et al., 2017) and that pedestrians may not even notice the absence of a driver (Rothenbücher et al., 2016). Moore et al. (2019) devoted an entire paper to arguing that eHMIs are superfluous, as pedestrians can judge whether it is safe to cross based solely on the kinematics of the approaching car.

In summary, research so far suggests that drivers' eye contact may encourage vulnerable road users to cross the road, but that implicit communication is more dominant. However, there appears to be neither systematic investigation that isolates drivers' eye contact from other confounding driver behaviors (e.g., being distracted) nor research about the effects of eye contact timing on pedestrians' perception of safety and crossing decisions. In the present online crowdsourced study, we examined participants' perceived safety to cross the road in front of an approaching

car, measured by means of a keypress response, for various timings of a driver's eye contact.

It was hypothesized that pedestrians feel safer to cross the road when the driver makes eye contact with them compared to when the driver does not make eye contact. The initiation of eye contact is a salient event due to the head turn involved. Thus, it is possible that not only eye contact itself, but also the *change* in the state of a driver's eye contact influences the pedestrian's crossing decisions. Furthermore, it is plausible that the highest perceived safety to cross is achieved when the change in a car's state of yielding (i.e., the initiation of braking) accompanies and complements a change in the state of a driver's eye contact (i.e., the initiation of eye contact).

2.2. Methods

2.2.1. Videos

Participants watched a set of 13 videos twice. Each video presented the viewpoint of a pedestrian standing on a sidewalk while a car (Smart Fortwo) with a driver was approaching from the left on a two-lane, 10 m-wide road. In 11 videos, the car yielded, and in 2 videos, it did not. Furthermore, in 11 videos, the driver made eye contact, and in 2 videos, he did not. The videos differed based on the initiation and termination of the driver's eye contact.

The videos were generated using an open-source simulator built in Unity3D (Bazilinskyy et al., 2020). They had a frame rate of 25 fps and a resolution of 1280×720 pixels. The videos included the engine sound of an approaching car (stereo, sample rate: 48 kHz). Videos were shown to participants via the cloud platform Heroku (https://www.heroku.com). The virtual camera in the animation was positioned 1.67 m above the pavement, which itself was 0.25 m above the road. The camera was also angled to obtain a full view of the road and was 0.7 m from the edge of the pavement. These values were regarded as comparable to the eye position of a typical pedestrian standing on the curb and turning to look at an approaching car. The field of view of the camera was set to relatively low values of 21 deg horizontally and 12 deg vertically, creating a 'zoomed in' effect. A narrow field of view was chosen to mimic the psychological experience of focused attention on approaching cars in real traffic. Wider fields of view were tried in the design of the experiment but were deemed less suitable, as in those cases, the car occupied a smaller part of the computer screen, which itself subtends only a limited field of view for the participant. Narrower fields of view, on the other hand, caused parts of the road and sidewalks to go out of sight, which was undesirable.

The 11 videos in which the car yielded were 31.0 s long, and the 2 videos in which the car did not yield were 21.0 s long. All videos started with a black screen lasting 1 s to prevent abrupt transitions between videos. The driver's eye contact in the videos was implemented by rotating the driver's head from its default straight-ahead

orientation to the orientation of the line connecting the driver's and pedestrian's heads. Initiation and termination of eye contact was achieved by turning the driver's head between these two orientations in 0.2–0.3 s. While making eye contact, the driver's head smoothly tracked the pedestrian's as the car approached.

The car's initial speed and longitudinal distance relative to the pedestrian (i.e., the camera's point of view in the videos) were 13.2 km/h and 66 m, respectively. It was expected that at lower speed, the effect of explicit communication, such as the driver's eye contact, becomes more important relative to the effect of implicit communication, such as the car's speed (Dey et al., 2021; Färber, 2016; Merat et al., 2018; Schneemann & Gohl, 2016).

In case the car yielded, it did so at a deceleration of 1 m/s^2, starting at a distance of 19.8 m (13.6 s) from the pedestrian and coming to a stop at a distance of 13.7 m (17.6 s). Distances were considered longitudinally between the location of the pedestrian's head (i.e., the camera's position) and the car's center. Given the Smart Fortwo's length of 2.695 m, the distance from the car's front end to the pedestrian at full stop was 12.35 m. Although the distance at stop was high compared to what one might see in real-life scenarios, it did not appear as high due to the low field of view (see Figure 2.1). Shorter distances were tried, but they caused the car to be partially or entirely out of view when stopped. The car stood still for 5.3 s, and then drove off with an acceleration of 1 m/s^2. Acceleration and deceleration values were set according to what was deemed a gentle change of speed for a Smart Fortwo, that is, about one-third of its maximum acceleration of 3.47 m/s^2, calculated based on a 0–60 km/h time of 4.8 s (Smart, 2021). The driver went out of sight 26.5 s into the videos involving a yielding car. For videos involving a non-yielding car, the driver went out of sight after 16.7 s from the start of the video.

Table 2.1 summarizes the characteristics of the videos in terms of yielding behavior of the car and eye contact interval. The reasoning behind the various eye contact intervals was that they represented all possible combinations between five distinct moments in a typical driver-pedestrian interaction:

1. The moment the approaching car is first visible ('First visible')
2. The moment the car starts to slow down ('Braking start')
3. The moment the car reaches a standstill in front of the pedestrian ('Full stop')
4. The moment the car starts to move again ('Take-off')
5. The moment the car leaves the pedestrian's view ('Out of sight')

In the case of a non-yielding car, only the first and last entries from the above are applicable. This rationale led to a total of 13 videos, each involving a unique interval of eye contact. Figure 2.1 showcases screenshots from videos with and without the driver's eye contact.

Table 2.1
Characteristics of the videos

Video	Yielding	Eye contact interval	Timing of eye contact
1	Yes	None	No eye contact
2	Yes	1.0–17.6	First visible–Full stop
3	Yes	1.0–22.9	First visible–Take-off
4	Yes	17.6–22.9	Full stop–Take-off
5	Yes	22.9–26.5	Take off–Out of sight
6	Yes	1.0–26.5	First visible–Out of sight
7	Yes	17.6–26.5	Full stop–Out of sight
8	Yes	1.0–13.6	First visible–Braking start
9	Yes	13.6–17.6	Braking start–Full stop
10	Yes	13.6–22.9	Braking start–Take-off
11	Yes	13.6–26.5	Braking start–Out of sight
12	No	None	No eye contact
13	No	1.0–16.7	First visible–Out of sight

Figure 2.1a. Screenshot from a video showing the driver making eye contact while the car was standing still.

Figure 2.1b. Screenshot from a video showing the driver not making eye contact while the car was standing still.

2.2.2. Participants

Two thousand participants were recruited from across the world via the online job portal Appen (https://www.appen.com). The job was titled "*Eye contact in traffic*". In Appen, participants first encountered a brief description of the study, followed by a question asking for informed consent. This research was approved by the Human Research Ethics Committee of the Delft University of Technology (reference number 1444).

2.2.3. Procedure

After providing informed consent, participants completed a questionnaire on their basic data and road behavior. Next, a link took them to the experiment on the Heroku platform, where the videos were preloaded to minimize delays, and the participants were presented with the following task instructions:

You will watch videos of approaching cars from the point of view of a pedestrian standing on the side of the road. Some cars will stop and other cars will not stop. In some videos, the driver will make eye contact with you. Imagine that you are the pedestrian and that you want to cross the road. Before the start of each video, you will briefly see a black screen. Please PRESS AND HOLD the 'F' key on your keyboard during this time. Once the video starts, continue holding the key as long as you feel safe to cross. RELEASE the key if you do not feel safe to cross anymore. You can press, hold and release the key as many times as you want per video.

Before proceeding, participants calibrated their device's volume against a piece of royalty-free music to ensure that they could hear the video sound clearly. They were then shown the 13 videos twice, all in random order, with a break after every 10 videos (the last batch contained only 6 videos). Under each video, the following text was present: "*PRESS AND HOLD the 'F' key when you feel safe to cross. RELEASE the key when you don't feel safe.*"

After each video, participants were presented with a question and a statement:

1) "*Did the driver make eye contact with you?*" (*No*, *Yes*)
2) "*The driver's eye contact behaviour was intuitive for me to decide whether I could or could not cross*" (five-point Likert scale ranging from '*Completely disagree*' to '*Completely agree*')

In addition, a total of 10 true-false test questions (e.g., "*Bananas are yellow*") were randomly inserted in each batch for each participant. These questions were used to screen out inattentive participants.

After all 26 videos were viewed, the penultimate page presented three additional statements:

1) "*Eye contact between drivers and pedestrians is important for road safety*"
2) "*I prefer eye contact to no eye contact*"
3) "*I could concentrate well during the study*"

Participants responded to these statements on five-point Likert scales of '*Completely disagree*' to '*Completely agree*'.

On the final page, participants received a unique worker code that they were required to enter in Appen as proof of completing the experiment and to receive payment. Each participant received a reimbursement of $0.45.

2.2.4. Analyses

First, participants who did not yield data or who may not have taken the task seriously were screened out. Next, trials that took longer than 33 s (for videos with a yielding car) or 23 s (for videos with a non-yielding car), i.e., more than 2 s longer than the nominal video duration were excluded. This exclusion was done to remove trials where participants may have suffered from technical problems such as lag while rendering in the browser or buffering of videos.

The first dependent variable, pedestrians' perceived safety, was analyzed by visualizing and statistically comparing the percentage of trials in which participants pressed the response key, for different videos. Statistical comparisons between videos were made at the level of participants using paired samples *t*-tests for each

0.1 s of video time. A conservative significance level (alpha = 0.001) was used to minimize the probability of false positives. Our approach is similar to Manhattan plots in molecular genetics, which use a stringent alpha value to visualize which of many genetic variants are predictive of a particular phenotype (Cook et al., 2013).

The second dependent variable was the mean intuitiveness rating of eye contact in helping make a crossing decision. Mean scores of the different videos were compared, and pairs of conditions were compared using paired samples *t*-tests, with an alpha value of 0.005 (Benjamin et al., 2018).

In addition, a performance score was calculated for the 11 videos depicting a yielding car. The performance score was calculated as: ((mean keypress percentage over the interval 13.6–22.9 s) + (100% − mean keypress percentage over the interval 22.9–26.5 s))/2. Accordingly, the performance score represents the extent to which participants felt safe to cross when it was indeed safe to cross (i.e., the braking and standing still phases) combined with the degree to which participants did not feel safe to cross when it was indeed unsafe to cross (i.e., the take-off phase).

2.3. Results

Participants who indicated that they did not read the instructions (*n* = 29), who indicated that they were younger than 18 (*n* = 3), who completed the study within 1000 s, suggesting cheating or carelessness (*n* = 89), who could not be linked to the data due to a data storage issue or cheating (*n* = 14), who made more than 2 mistakes out of the 10 test questions (*n* = 16), or who suffered video playback delays, defined as more than 2 videos taking more than 5 s too long to complete (*n* = 49) were excluded, leaving 1835 participants from 64 countries. Multiple participations from the same IP address were permitted, as there was no reliable way to determine whether duplicate IPs were due to one person or multiple persons completing the experiment on one device or multiple devices connected to the same network. Out of the total number of 47710 trials (1835 participants x 26 trials per person), 46277 trials (97.0%) were retained, whereas the rest of the trials were excluded due to playback lags of more than 2 s.

The mean study completion time was 39.1 min (*SD* = 18.8 min, median = 33.0 min). The study yielded a mean satisfaction score of 4.4 on a scale of 1 (*very dissatisfied*) to 5 (*very satisfied*) by 86 people who completed the optional satisfaction survey offered by Appen. The five most represented countries were Venezuela (*n* = 1098), the United States (*n* = 210), Russia (*n* = 70), India (*n* = 59), and Egypt (*n* = 57). The participants consisted of 1159 males, 668 females, and 8 people who indicated ‘I prefer not to respond’. The mean age was 34.9 years (*SD* = 10.9). A total of 100 participants indicated that they were ‘never’ pedestrians, 68 indicated ‘less than once a month’, 166 ‘once a month to once a week’, 487 ‘1 to 3 days a week’, 376 ‘4 to 6 days a week’, and 580 ‘every day’ (58 participants indicated ‘I prefer not to respond’).

2.3.1. Keypress Percentage as a Measure of Perceived Safety to Cross

Figure 2.2 shows the mean keypress percentage for yielding cars for driver's eye contact throughout versus no driver's eye contact. It can be seen that pedestrians perceived the situation as less and less safe until the car began braking. As the car braked, perceived safety increased and remained high while the car was fully stopped, only to drop sharply after the car took off. This sharp drop started at 200–300 ms and has the highest slope at 300–600 ms after the car drove off. This drop is consistent with the reaction time distribution to a discrete stimulus (e.g., Ratcliff, 1993), which in this case would be the onset of motion of the car. Figure 2.2 further shows that there was also a small increase in the mean keypress percentage after the car went out of sight.

It can also be seen from Figure 2.2 that the driver's eye contact did not significantly affect perceived safety before the car started braking and for most of the car's take-off. The driver's eye contact substantially increased perceived safety compared to a lack of it in a time window starting soon after the car started braking and ending shortly after the car took off again.

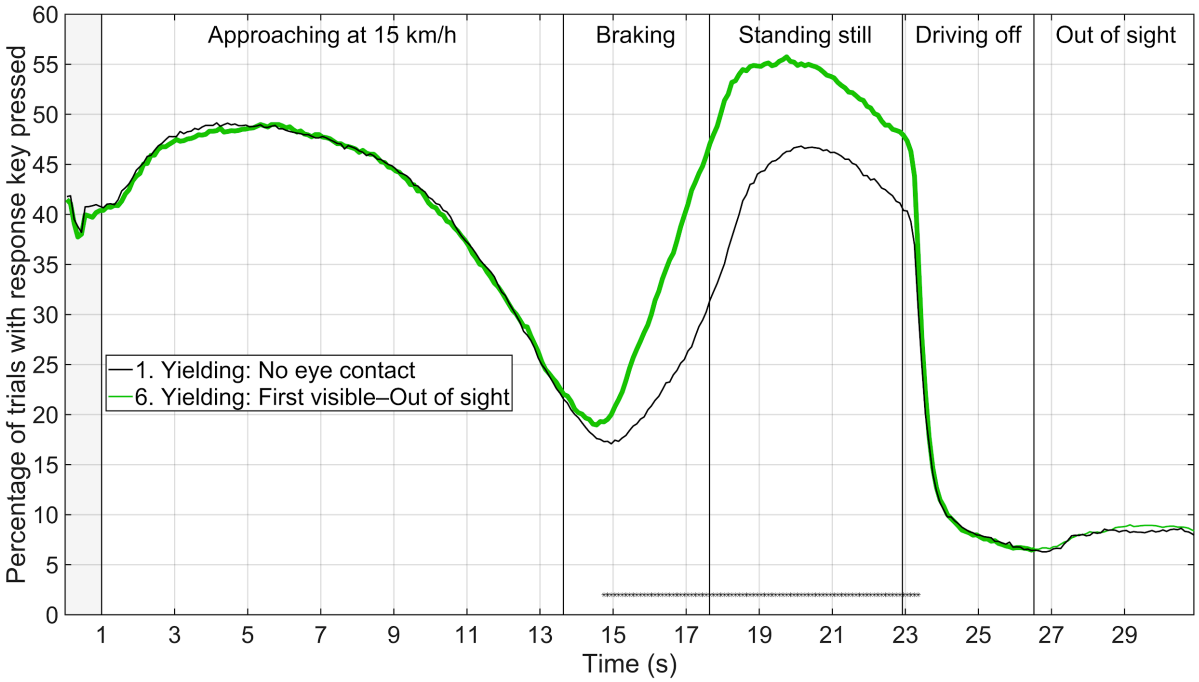

Figure 2.2. Percentage of trials in which the response key was pressed for no driver's eye contact (Video 1) and driver's eye contact throughout (Video 6), for the videos in which the car yielded. The bold sections of the lines indicate that there was eye contact at those moments. The asterisks at the bottom indicate significant differences, $p < 0.001$.

For non-yielding cars (Figure 2.3), perceived safety decreased throughout the video, which is explained by the fact that the car got closer and closer to the pedestrian but without slowing down. Similar to Figure 2.2, a slight increase in perceived safety can be seen after the non-yielding car went out of sight after passing the pedestrian.

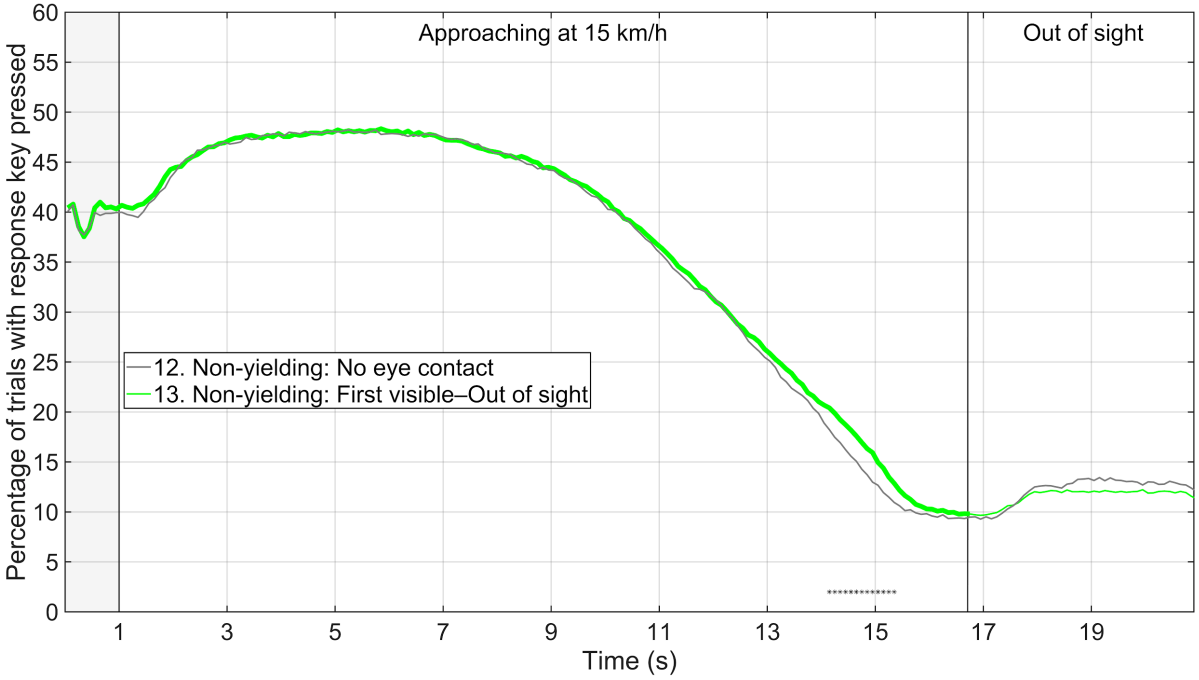

Figure 2.3. Percentage of trials in which the response key was pressed for no driver's eye contact (Video 12) and driver's eye contact throughout (Video 13), for the videos in which the car did not yield. The bold sections of the lines indicate that there was eye contact at those moments. The asterisks at the bottom indicate significant differences between pairs of conditions, $p < 0.001$.

Figure 2.4 shows that if the driver started making eye contact when the car started to brake (Video 11), perceived safety during the braking phase was significantly higher compared to when eye contact was already present at the beginning of the video (Video 6). In other words, the initiation of eye contact alongside the initiation of braking positively affected pedestrians' perceived safety compared to eye contact throughout the encounter. From Figure 2.4, it can be seen that eye contact while braking had a strong effect: when the car came to a stop, the keypress percentage was 31.3% for Video 7 but 49.7% for Video 11 (i.e., a 59% increase).

Figure 2.4 further shows that the initiation of eye contact when the car came to a stop (Video 7) gave a small boost to perceived safety, so that in parts of the standing still and driving off phases, it was higher compared to eye contact throughout (Video 6). So again, the initiation of eye contact had a reinforcing effect on the keypress percentage compared to continuous eye contact from the beginning of the video.

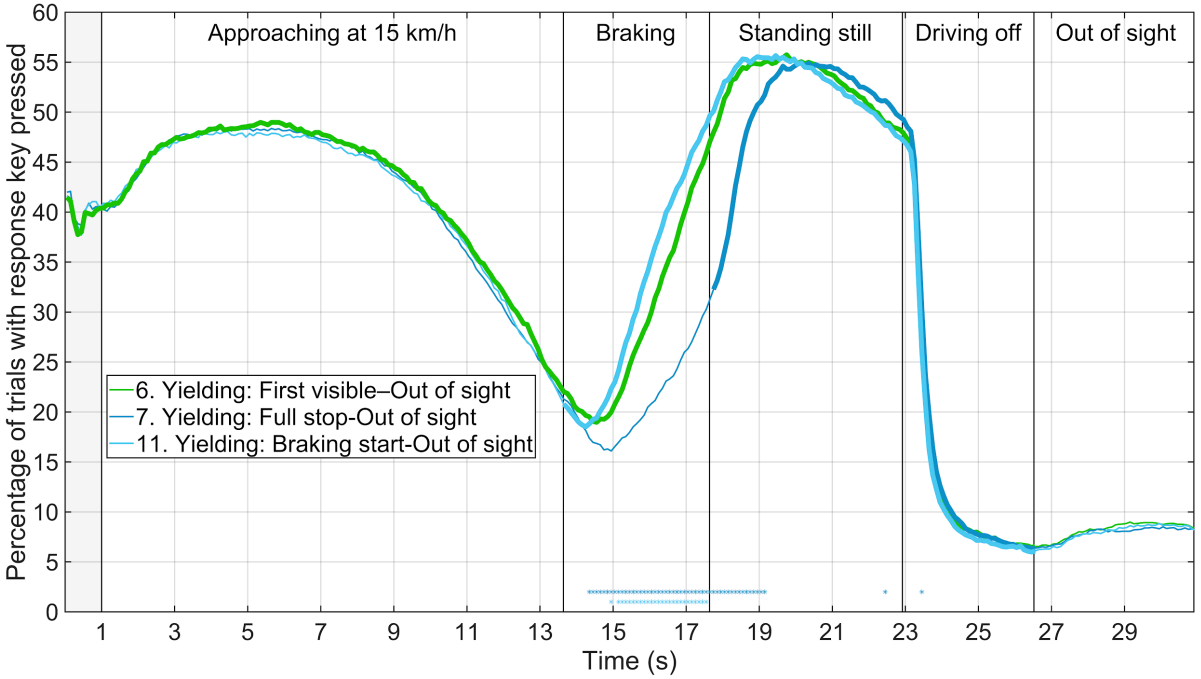

Figure 2.4. Percentage of trials in which the response key was pressed for driver's eye contact throughout (Video 6), and eye contact initiation when the car started to brake (Video 11), and when the car came to a stop (Video 7). The bold sections of the lines indicate that there was eye contact at those moments. The asterisks at the bottom indicate significant differences with the video containing eye contact throughout (Video 6), $p < 0.001$.

If the driver stopped making eye contact when the car came to a full stop (Videos 2 and 9), this was seen by pedestrians as a sign that they should not cross when the car was standing still, compared to no eye contact at all (Video 1), as seen in Figure 2.5. In other words, just like the initiation of eye contact was a cue that pedestrians should cross, the termination of eye contact was a cue that they should not cross. It is interesting that the effects of eye contact termination at full stop carried forward until after take-off. It is also worth noting that the drop in mean keypress response due to termination of eye contact was not as steep as the drop due to the car's initial approach or its take-off, suggesting that the car's motion was a more dominant cue.

The above figures show that participants' perceived safety reduced abruptly when the car started to drive away, regardless of eye contact. In other words, implicit communication (i.e., vehicle motion) was more dominant. There was, however, a delayed response for cases when the driver retained eye contact while driving away (e.g., Video 6) compared to when eye contact ended when the car drove off (e.g., Video 3), as seen by the presence of significant differences in Figure 2.6. Thus again, the termination of eye contact was perceived as a sign that the pedestrian should not cross.

Figures 2.2–2.6 showed results for selected videos. The results for all 11 videos involving a yielding car are available in Appendix 2.A (Figure 2.A1).

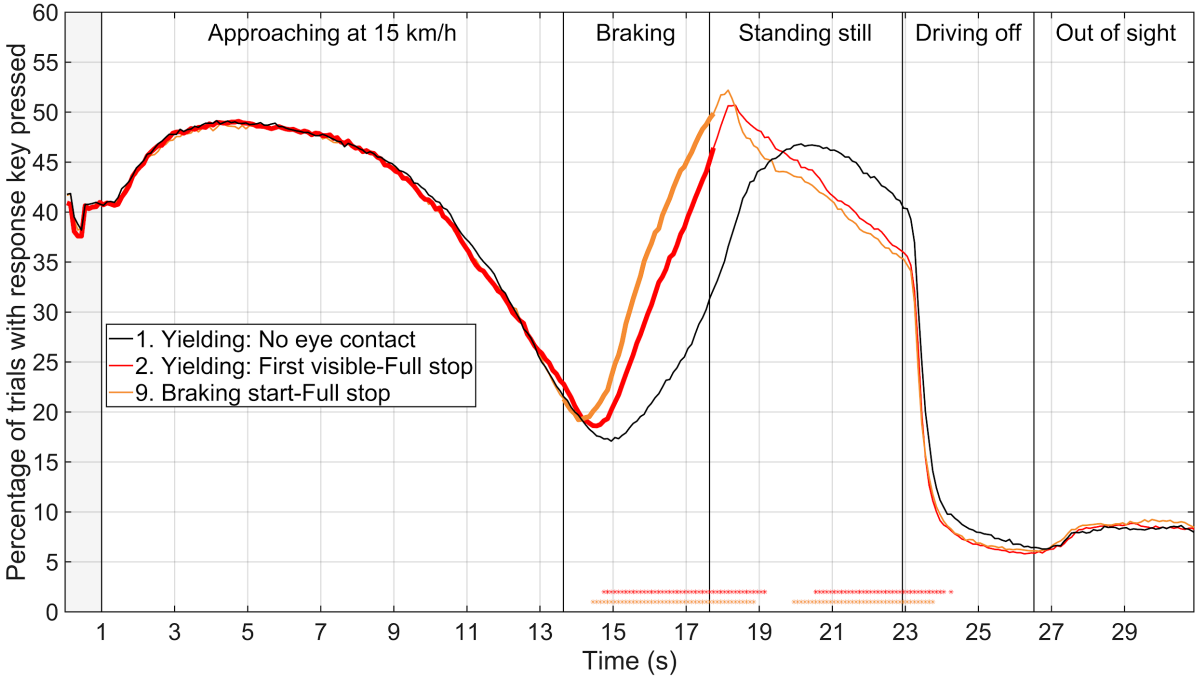

Figure 2.5. Percentage of trials in which the response key was pressed for no driver's eye contact throughout (Video 1) and eye contact termination when the car came to a stop (Videos 2 and 9). The bold sections of the lines indicate that there was eye contact at those moments. The asterisks at the bottom indicate significant differences with the video containing no eye contact (Video 1), $p < 0.001$.

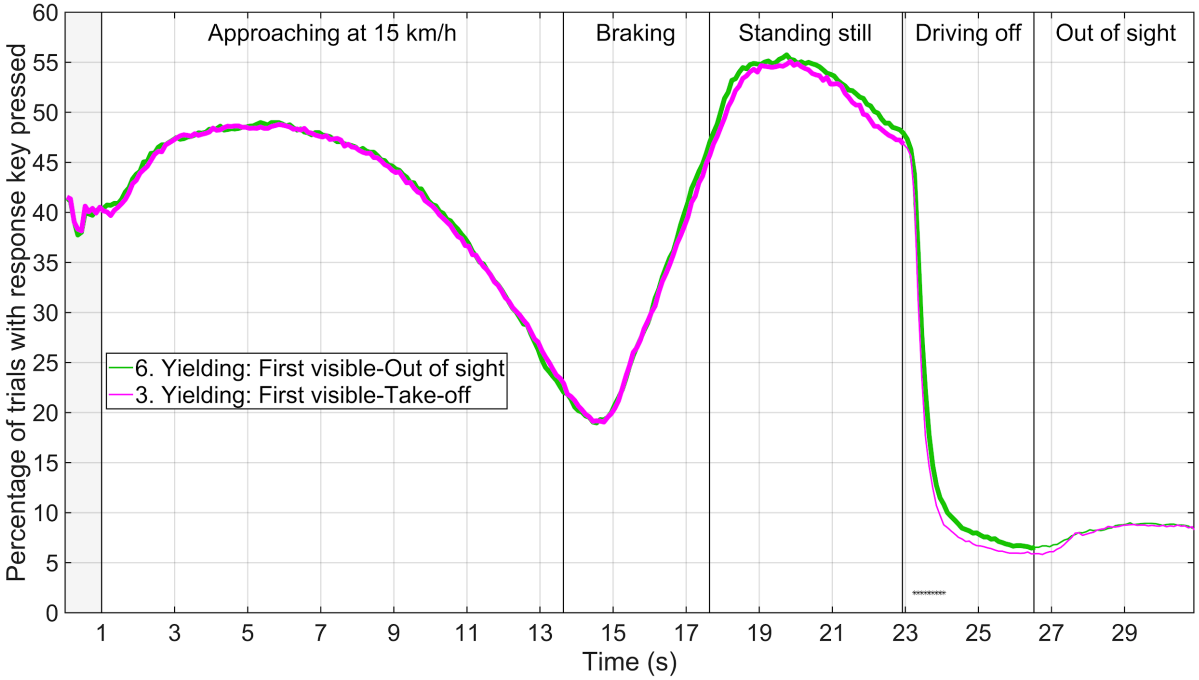

Figure 2.6. Percentage of trials in which the response key was pressed for driver's eye contact throughout (Video 6) and eye contact termination when the car took off (Video 3). The bold sections of the lines indicate that there was eye contact at those moments. The asterisks at the bottom indicate significant differences, $p < 0.001$.

2.3.2. Relationships Between Self-Reported Intuitiveness, Self-Reported Driver's Eye Contact, and Objective Performance

Figure 2.7 provides an indication of concurrent validity by depicting the mean scores across participants for self-reported intuitiveness, self-reported driver's eye contact, and objective performance. From Figure 2.7 (left), it can be seen that for videos in which there was eye contact, self-reported eye contact was high (i.e., close to 100%), suggesting that participants were generally attentive.

Figure 2.7 (left) also shows that the false-positive rate was 15–20% for the videos without eye contact (Videos 1 and 12). These values are relatively far from 0%, which can be explained by the fact that it is difficult to ascertain that there was no eye contact at all during the entire video because of the low signal-to-noise ratio when the car is still far away. That is, when the car is far away, participants have to guess whether there is eye contact or not since they cannot see the driver clearly.

Intermediate percentages of self-reported eye contact were observed for Video 5 (85%), Video 8 (38%), and Video 13 (84%). These relatively low percentages may be because eye contact occurred late in the video, only when the car drove off (Video 5), occurred very early, in which case eye contact is hard to detect (Video 8), or because the car did not stop, which may have also made it difficult to detect eye contact (Video 13).

Figure 2.7 (left) further shows clear differences between the intuitiveness ratings of the 13 videos. From the 78 pairs of comparisons between the 13 videos, only 7 pairs were not significantly different from each other, i.e., p greater than 0.005 (Video pairs 1–12, 2–7, 3–10, 5–8, 6–9, 6–11, and 9–11), indicating that our study was adequately powered to detect small differences in the intuitiveness ratings. The highest intuitiveness ratings were found for Videos 3 and 10, which were videos in which the driver terminated eye contact upon driving away, whereas the highest performance scores were obtained for Videos 10 and 11 (Figure 2.7, right), which were videos in which the driver initiated eye contact when the car started braking.

Among the videos in which the driver made eye contact, Video 5 was the least intuitive and yielded the lowest performance by a substantial margin (Figure 2.7, right). In this video, the driver started looking at the pedestrian when the car started to drive off. Another counterintuitive video that yielded low performance was Video 7. This video was similar to Video 5 as the driver started to make eye contact upon take-off.

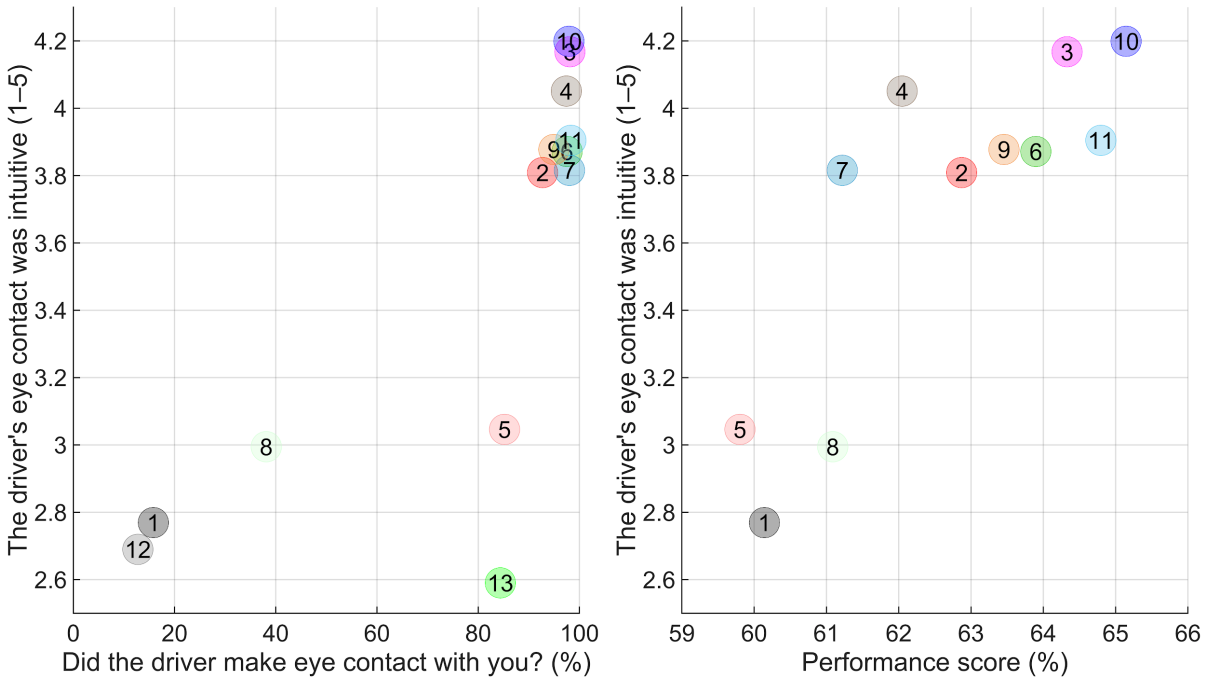

Figure 2.7. Mean self-reported intuitiveness of eye contact per video versus self-reported occurrence of eye contact per video (left), and mean self-reported intuitiveness of eye contact per video versus performance score per video with a yielding car (right).

2.3.3. Effects of Self-Reported Concentration

An issue in online studies such as the present one is that inattentive participants may contaminate the data. A small positive correlation was observed between self-reported concentration and objective performance (r = 0.10, p < 0.001, n = 1813 participants with a response to this question). The association between self-reported concentration and keypress behavior for videos with yielding cars is illustrated in Figure 2.8, showing that non-concentrated participants were less likely to hold the key than concentrated participants. More specifically, the mean keypress percentages from the start of the video until the car went out of sight were 20, 31, 32, 37, and 37% for concentration levels 1 (*Completely disagree*) to 5 (*Completely agree*).

2.3.4. Cross-Cultural Consistency

A common question in the analysis of eye contact and other road user gestures is whether there may exist cross-cultural differences (Ranasinghe et al., 2020). In an attempt to address this question, we computed the means and standard deviations, as well as correlations of the means of videos (n = 13) for participants from the five most represented countries.

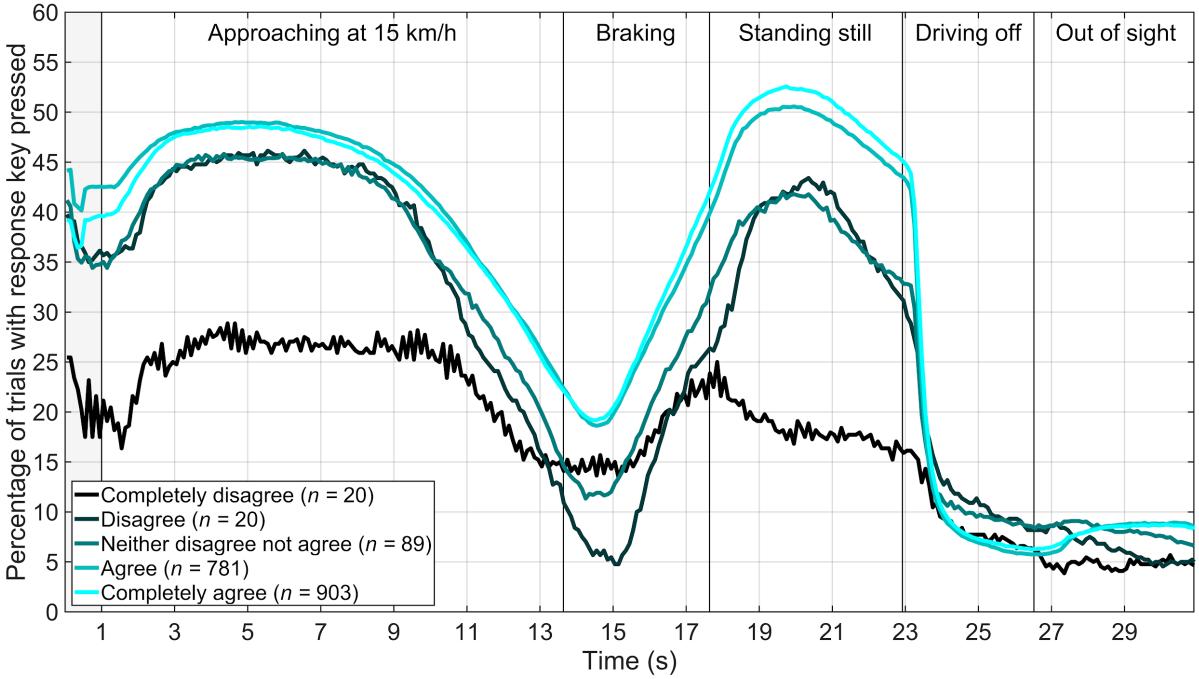

Figure 2.8. Percentage of trials in which the response key was pressed for different self-reported concentration levels.

The results in Table 2.2 suggest that outcomes for participants from different countries were highly similar. More specifically, the average ratings of driver's eye contact intuitiveness were similar (between 3.44 for Egypt and 3.56 for Venezuela, on the scale of 1–5), and the average performance scores were similar as well (between 61% for Egypt and 63% for the United States). Also, the correlations of the mean intuitiveness ratings were all high ($r > 0.97$, $n = 13$) and correlations for the mean performance scores were high as well ($r > 0.90$, $n = 11$), with the exception of participants from Egypt, whose performance scores showed a more modest correlation with the performance scores of participants from the four other countries. Nonetheless, correlations between intuitiveness ratings and performance scores were all around $r = 0.75$ ($n = 11$), indicating that the results presented in Figure 2.7 (right) are cross-nationally robust. The correlation coefficients for the two most highly represented countries (Venezuela and the United States) are illustrated in Figure 2.9.

The performance scores were computed based on whether the participants pressed the key when the key should be pressed and released the key when the key should not be pressed. Although performance scores were similar, the base rates of key presses were different between countries, with mean keypress percentages from the start of the video until the car went out of sight being 36, 33, 47, 39, and 33% for Venezuela, United States, Russia, India, and Egypt, respectively. These differences in base rates, which are illustrated in Appendix 2.A (Figure 2.A2), may be caused by some participants from particular countries misunderstanding the task or not taking the task seriously (e.g., holding the key throughout the trial). Such anomalies were,

however, not of concern for the relative effects between videos, as was demonstrated in Table 2.2.

Table 2.2
Means, standard deviations, and Pearson product-moment correlation coefficients of means per video (n = 13 for intuitiveness ratings, and n = 11 for performance scores, which were computed for videos with a yielding car) for participants from different countries

		M	*SD*	1	2	3	4	5	6	7	8	9	10	11
1	Intuitive (1–5) (All)	3.52	0.60											
2	Intuitive (1–5) (VEN)	3.56	0.62	0.999										
3	Intuitive (1–5) (USA)	3.53	0.54	0.998	0.999									
4	Intuitive (1–5) (RUS)	3.47	0.78	0.987	0.981	0.978								
5	Intuitive (1–5) (IND)	3.46	0.54	0.981	0.978	0.971	0.978							
6	Intuitive (1–5) (EGY)	3.44	0.50	0.987	0.987	0.981	0.972	0.979						
7	Performance (%) (All)	62.62	1.87	0.827	0.823	0.830	0.805	0.760	0.830					
8	Performance (%) (VEN)	62.71	1.74	0.834	0.832	0.836	0.803	0.772	0.837	0.997				
9	Performance (%) (USA)	62.84	1.70	0.776	0.768	0.777	0.771	0.693	0.789	0.980	0.964			
10	Performance (%) (RUS)	61.45	3.22	0.871	0.864	0.875	0.869	0.800	0.861	0.963	0.950	0.957		
11	Performance (%) (IND)	61.42	2.05	0.788	0.779	0.786	0.795	0.725	0.807	0.925	0.905	0.959	0.915	
12	Performance (%) (EGY)	60.75	1.70	0.749	0.748	0.739	0.703	0.731	0.776	0.866	0.884	0.814	0.844	0.729

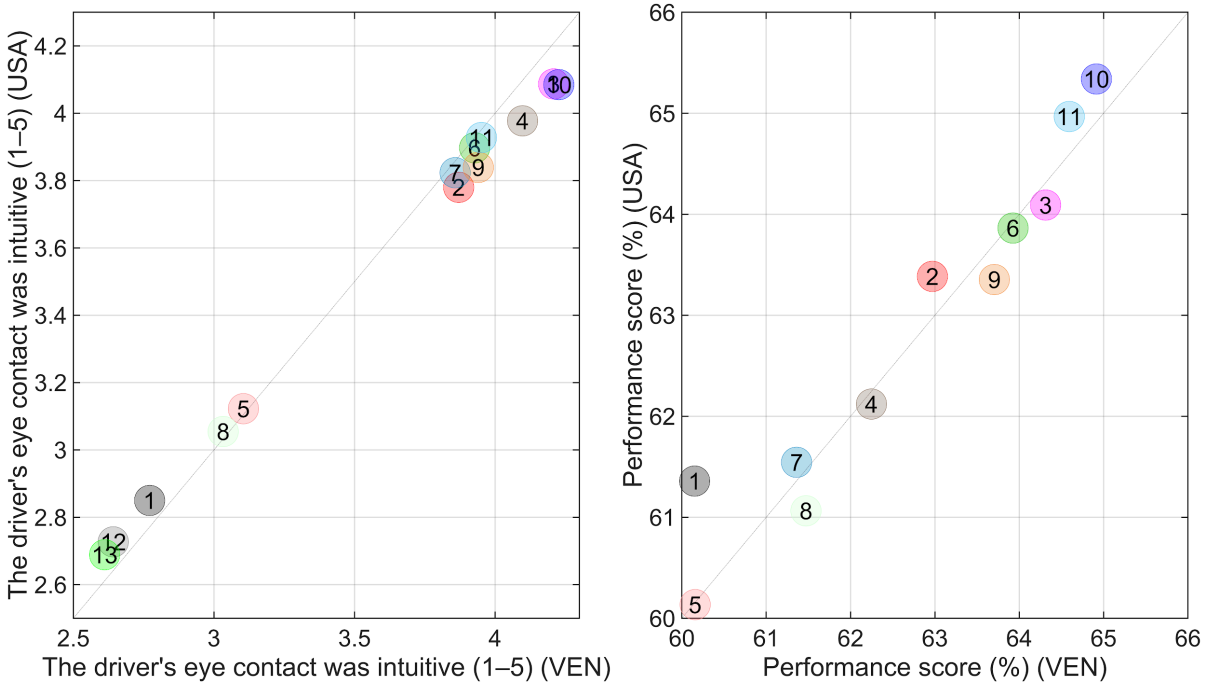

Figure 2.9. Mean self-reported intuitiveness of eye contact per video for participants from the United States and Venezuela (left, *r* = 0.999), and mean performance score per video involving a yielding car for participants from the United States and Venezuela (right, *r* = 0.964). The diagonal lines are lines of unity.

2.4. Discussion

This study aimed to examine the effect of drivers' eye contact on pedestrians' perceived safety to cross a road. Through an online experiment with a large sample size, we varied the start and end times of a driver's eye contact, yielding a total of ten different eye contact intervals for an approaching car that slowed down to a full stop and subsequently drove off. In addition, we included a video with no eye contact as a baseline and two videos in which the approaching car did not yield: one with eye contact throughout and one without eye contact.

The results of this study unambiguously indicated that a driver's eye contact made pedestrians feel safer to cross. This finding confirms the scarce evidence so far that suggests that eye contact increases the feeling of safety and willingness to cross (Bazilinskyy et al., 2023; Malmsten Lundgren et al., 2017; Yang, 2017). The studies so far, however, did not provide insight into the effects of drivers' eye contact on pedestrians as a function of time during the car's approach.

In our study, the effects of different eye contact timings were investigated, and the results can be summarized by stating that not only eye contact but also the initiation and termination of eye contact affect perceived safety. That is, the initiation of eye contact alongside braking was found to be a more powerful cue for pedestrians to cross compared to eye contact throughout, and conversely, the termination of eye contact alongside take-off was a stronger deterrent to cross than no eye contact at all. This argument may be extended to say that there exists a time window between a car's braking and its subsequent take-off where eye contact is a strong cue to help resolve crossing conflicts.

Previous research made a case against the importance of eye contact and the usefulness of eHMIs, by arguing that implicit communication alone is sufficient for pedestrians, without any need for explicit communication, such as eye contact (Moore et al., 2019). The present study does not dispute that a car's motion is a more dominant cue than eye contact; in fact, it confirms this. However, it also provides counterevidence to the claim that eye contact is dispensable by showing that eye contact initiation while braking increased the perceived safety, with an increase from 31% to 50% of participants feeling safe to cross the road when the car came to a stop. These findings have implications for research into substituting eye contact in the context of AVs, i.e., that replacements may indeed be required to maintain the safety of pedestrians.

As pointed out above, our study showed that implicit communication could override the effect of eye contact. For example, the driver's eye contact did not have much of an effect if the car did not slow down. This finding can be explained by the fact that crossing will lead to collision in this scenario and is therefore intuitively unsafe. Similarly, after the car drove off from a standstill, participants consistently released the response key, and eye contact had a comparatively small effect. These results

can be summarized by the common-sense notion that eye contact, although a compelling cue, is not compelling enough to cause participants to get run over by the car. Our study further showed that eye contact failed to have an effect when eye contact could not be detected with certainty, that is, when the vehicle was still far away. If the driver's eyes or head movement cannot be seen because of the large distance, pedestrian behavior cannot be affected.

The current study investigated the effect of drivers' eye contact on pedestrians' perceived safety. The converse topic, namely the effect of pedestrians' eye contact on drivers' perceived safety, would be of interest as well. Several early studies showed that drivers slowed down or stopped more often when staged pedestrians/hitchhikers looked at the approaching vehicle compared to when they did not (Katz et al., 1975; Morgan et al., 1975; Snyder et al., 1974). Similarly, Ren et al. (2016) observed that drivers braked earlier and approached more slowly when staged pedestrians made eye contact with them as opposed to when they did not. Naturalistic driving studies suggest that pedestrians' gaze/eye contact combined with other pedestrian behaviors such as facial expression and assertiveness have important roles in successfully resolving driver-pedestrian interactions (Kong et al., 2021; Nathanael et al., 2019; Uttley et al., 2020). The relatively small number of studies so far suggests that more research is needed in the area of the effect of pedestrians' eye contact on drivers. Apart from investigating one-way communication (i.e., driver→pedestrian, pedestrian→driver), it would be worthwhile to examine reciprocal effects of eye contact on both drivers and pedestrians, taking into consideration that eye contact is both an input (i.e., reading the other agent's intentions) and a cue (i.e., signaling one's own intentions), cf. Myllyneva and Hietanen (2016). The notion of mutual attention in traffic is a topic that is receiving increased attention nowadays (Kotseruba et al., 2016; Onkhar et al., 2021).

Some limitations have to be acknowledged. In particular, participants were looking at a monitor, were not immersed in actual traffic, and did not experience physical risk. In real traffic, pedestrians might overlook drivers' eye contact or have particular incentives to cross the road, for example, being in a hurry (Cefkin et al., 2019). The detectability of eye contact in our study may be better or worse than the detectability of eye contact in real traffic. In our study, participants watched videos with a resolution of 1280×720 pixels. Based on a side-by-side comparison of video frames, eye contact (i.e., head turn) was already distinguishable from no eye contact when the car was about 50 m away. However, when eye contact was present from the video start until a 20 m distance, only 38% of participants reported noticing it (see Video 8 in Figure 2.7, left), which is modestly higher than the video without eye contact (Video 1, with 16% of participants reporting eye contact). These findings are supported by Figure 2.A3 in Appendix 2.A, showing the percentage of trials in which the response key was pressed as a function of vehicle-pedestrian distance. It can be seen that the earliest deviation between eye contact from the video start and no eye contact arose at a distance of about 25 m. In other words, although eye contact may

have been theoretically detectable at farther distances, in our experiment, pedestrians started noticing and reacting to the driver's eye contact at a vehicle-pedestrian distance of 20–25 m or less. Research in real-life conditions shows that the detectability of eye contact (Martin & Jones, 1982; examined distances between 0.6 and 4.0 m), facial affect (Hager & Ekman, 1979; examined distances between 30 m and 45 m) and the recognizability of individuals (Lampinen et al., 2014; examined distances between 3.7 and 37 m) decreases with increasing distance. In the real world, pedestrians are not hampered by restrictions of screen resolution, but other factors may impair the detectability of eye contact at a distance, such as smog, blinding headlamps at night, shadows, and windshield glare (e.g., AlAdawy et al., 2019; Schneemann & Gohl, 2016). Thus, it remains to be investigated how well the present findings are applicable to actual on-road settings.

It is to be noted that the speed of the car was 13 km/h, representative of speeds in residential areas and shared spaces. At this low speed, there is presumably substantial uncertainty about what the car will do. It can be expected, based on related findings in the literature (Schneemann & Gohl, 2016), that if the car would approach at higher speeds, then perceived safety would be lower, and the effect of eye contact relative to implicit communication would be smaller or effective across a shorter time interval. It needs to be investigated how the results generalize to high-speed interactions.

It should also be noted that all conditions included a male driver with a presumed neutral or somewhat authoritative expression on his face. It would be relevant to investigate whether the results apply to different types of drivers as well. Previous research indicates that emotional expression (e.g., happy, sad, angry) has a strong effect on perceived dominance (Sutton et al., 2019) and may be interpreted differently cross-culturally (Arapova, 2017). Another limitation is that most of the participants were from Venezuela, followed by the United States, which may be regarded as an idiosyncratic subset of the world population. While the current study showed that the effects of drivers' eye contact generalize well between participants from different countries (see also Bazilinskyy et al., 2023), cultural differences in vehicle-pedestrian interactions may still exist. Norman (1992) anecdotally reflected on eye contact in Mexico City traffic: "*it was essential to avoid eye contact with other drivers. In the traffic circles of the city, the trick was to avoid letting the other drivers see that you had seen them. Once the other drivers knew that you knew they were there, they would proceed at high speed around the circle, completely ignoring your presence, because they knew that you knew that they were there, so they expected you to stop or slow down. … Most places in the United States don't let you get away with such games.*" (also see Vanderbilt, 2008, making the same point about driving in Mexico City). It would be interesting to explore these and other cultural differences in future research.

In relation to the above, some eye contact behaviors by the driver in our videos may be perceived as unnatural. Continuous eye contact by the driver throughout the car's approach is one such example. However, in the interest of systematically varying eye contact across the videos, it was necessary to include this scenario. Additionally, some eye contact behaviors by the driver that one might find realistic were deliberately excluded in our videos, such as multiple back-and-forth looks by the driver during the car's approach. These were undesirable in our experiment's design as they introduced a confounding variable – the number of eye contact attempts in a single interaction, which would have taken the focus away from eye contact duration and initiation/termination.

Another point of attention is that some participants may not have concentrated on the task or may have misunderstood the task. In particular, the results showed that about 8% of the participants inappropriately held the key when it was unsafe to cross, that is, when the vehicle drove away. The test questions and self-reports, on the other hand, revealed committed participants, with only 16 of 2000 participants failing the test questions (more than 2 mistakes out of 10 questions), and correct detection rates of eye contact close to 100% for videos in which there was eye contact when the car was close. Even though not all participants were fully committed to the task, this should not affect relative comparisons between the results for the different videos.

Finally, while not a shortcoming per se, it is worth noting that the current study is another entry in a series of papers we have published on pedestrians' crossing behavior that employed online crowdsourcing and our open-source simulator (Oudshoorn et al., 2021; Sripada et al., 2021). As such, this paper bears similarities in its methods with its predecessors but also retains its novelty as an attempt to further the understanding of eye contact in traffic, since the prior works were instead concerned with eHMIs and vehicle motion.

In conclusion, this experiment demonstrated for the first time how drivers' eye contact and its timing affect the perceived safety of pedestrians. Results indicate that the presence and initiation of eye contact increase perceived safety, whereas the absence and termination of eye contact reduce perceived safety. The results also suggest that eye contact helps resolve crossing conflicts during a time window starting from the car's braking and ending with its subsequent take-off, and that a replacement for eye contact may be needed in the context of AVs. Future research could repeat the present study in a staged on-road design. Future research could also examine whether the driver's eye rotation or the driver's head rotation is a more dominant factor in pedestrians' perceived safety.

Acknowledgments

This research is funded by grant 016.Vidi.178.047 ("How should automated vehicles communicate with other road users?"), which is provided by the Netherlands

Organization for Scientific Research (NWO). We would like to thank Piotr Sienkowski for his help with modifying the simulator.

Appendix 2.A.

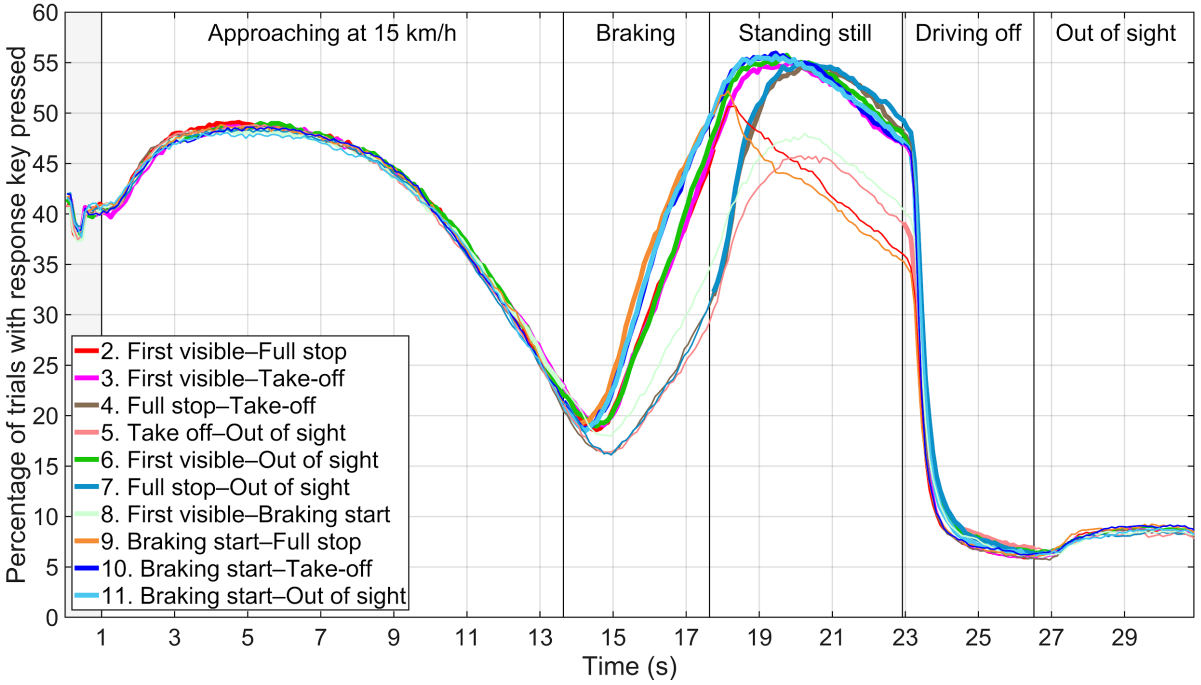

Figure 2.A1. Percentage of trials in which the response key was pressed for the videos that depict a yielding car. The bold sections of the lines indicate that there was eye contact at those moments.

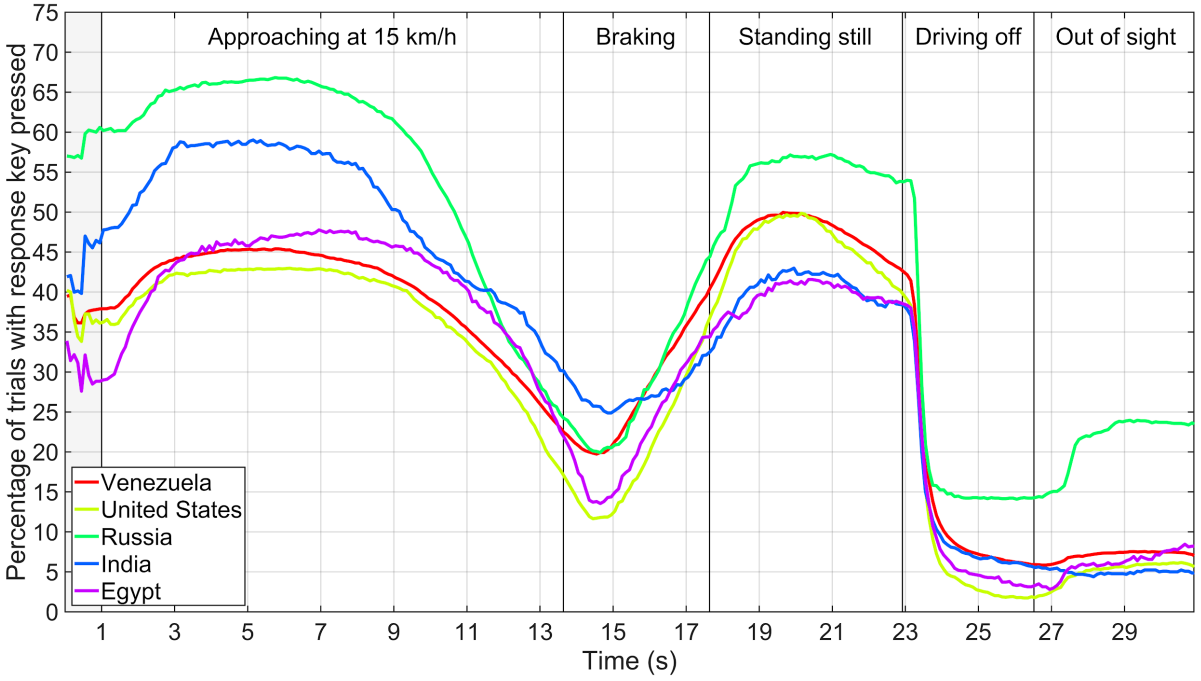

Figure 2.A2. Percentage of trials in which the response key was pressed for the videos that depict a yielding car for participants from different countries. The responses for the 11 videos were averaged.

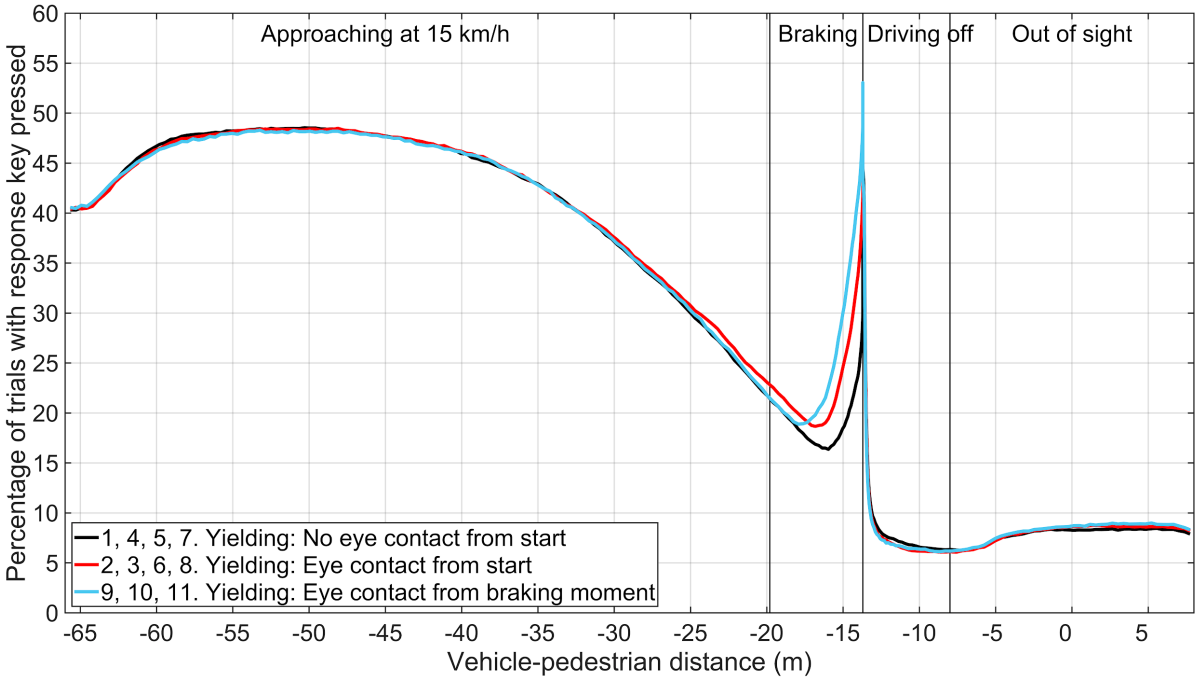

Figure 2.A3. Percentage of trials in which the response key was pressed for no driver's eye contact from the start of the video (Videos 1, 4, 5, and 7 averaged), driver's eye contact from the start of the video (Videos 2, 3, 6, and 8 averaged), and eye contact initiation when the car started to brake (Videos 9, 10, and 11 averaged) vs. vehicle-pedestrian distance.

Supplementary Data

The videos shown, questions posed, data collected, and MATLAB code used for the analysis are available at: https://doi.org/10.4121/16866709.

References

AlAdawy, D., Glazer, M., Terwilliger, J., Schmidt, H., Domeyer, J., Mehler, B., Reimer, B., & Fridman, L. (2019). Eye contact between pedestrians and drivers. *Proceedings of the Tenth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Santa Fe, New Mexico, 301–307. https://pubs.lib.uiowa.edu/driving/article/28343/galley/136635/view

Alvarez, W. M., De Miguel, M. Á., García, F., & Olaverri-Monreal, C. (2019). Response of vulnerable road users to visual information from autonomous vehicles in shared spaces. *IEEE Intelligent Transportation Systems Conference*, Auckland, New Zealand, 3714–3719. https://doi.org/10.1109/ITSC.2019.8917501

Alvarez, W. M., Moreno, F. M., Sipele, O., Smirnov, N., & Olaverri-Monreal, C. (2020). Autonomous driving: Framework for pedestrian intention estimation in a real world scenario. *IEEE Intelligent Vehicles Symposium*, Las Vegas, NV, 39–44. https://doi.org/10.1109/IV47402.2020.9304624

Arapova, M. A. (2017). Cultural differences in Russian and Western smiling. *Russian Journal of Communication, 9*, 34–52. https://doi.org/10.1080/19409419.2016.1262208

Argyle, M., & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, *28*, 289–304. https://doi.org/10.2307/2786027

Bazilinskyy, P., Dodou, D., Eisma, Y. B., Vlakveld, W., & De Winter, J. C. F. (2023). Blinded windows and empty driver seats: The effects of automated vehicle characteristics on cyclists' decision-making. *IET Intelligent Transportation Systems*, *17*, 72–84. https://doi.org/10.1049/itr2.12235

Bazilinskyy, P., Kooijman, L., Dodou, D., & De Winter, J. C. F. (2020). Coupled simulator for research on the interaction between pedestrians and (automated) vehicles. *Proceedings of the Driving Simulation Conference Europe*, Antibes, France. https://repository.tudelft.nl/islandora/object/uuid:e14ae256-318d-4889-adba-b0ba1efcca71

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Cefkin, M., Zhang, J., Stayton, E., & Vinkhuyzen, E. (2019). Multi-methods research to examine external HMI for highly automated vehicles. In H. Krömker (Ed.), *HCI in Mobility, Transport, and Automotive Systems* (pp. 46–64). Springer. https://doi.org/10.1007/978-3-030-22666-4_4

Chang, C.-M., Toda, K., Sakamoto, D., & Igarashi, T. (2017). Eyes on a car: An interface design for communication between an autonomous car and a pedestrian. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Oldenburg, Germany, 65–73. https://doi.org/10.1145/3122986.3122989

Clamann, M., Aubert, M., & Cummings, M. L. (2017). Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles. *Transportation Research Board 96th Annual Meeting*, Washington DC, Article 17-02119. https://www.researchgate.net/publication/312155228_Evaluation_of_Vehicle-to-Pedestrian_Communication_Displays_for_Autonomous_Vehicles

Cook, D. E., Ryckman, K. R., & Murray, J. C. (2013). Generating Manhattan plots in Stata. *The Stata Journal*, *13*, 323–328. https://doi.org/10.1177/1536867X1301300206

Dey, D., Matviienko, A., Berger, M., Pfleging, B., Martens, M., & Terken, J. (2021). Communicating the intention of an automated vehicle to pedestrians: The contributions of eHMI and vehicle behavior. *it-Information Technology, 63*, 123–141. https://doi.org/10.1515/itit-2020-0025

Dey, D., & Terken, J. (2017). Pedestrian interaction with vehicles: Roles of explicit and implicit communication. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Oldenburg, Germany, 109–113. https://doi.org/10.1145/3122986.3123009

European Road Safety Observatory. (2018). Traffic safety basic facts 2018. Pedestrians. https://road-safety.transport.ec.europa.eu/document/download/0d44bd4f-7e35-4af5-86ce-2675f6339c77_en?filename=bfs2018_main_figures.pdf

Faas, S. M., Stange, V., & Baumann, M. (2021). Self-driving vehicles and pedestrian interaction: Does an external human-machine interface mitigate the threat of a tinted windshield or a distracted driver? *International Journal of Human–Computer Interaction, 37*, 1364–1374. https://doi.org/10.1080/10447318.2021.1886483

Färber, B. (2016). Communication and communication problems between autonomous vehicles and human drivers. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous driving* (pp. 125–144). Springer. https://doi.org/10.1007/978-3-662-48847-8_7

Furuya, H., Kim, K., Bruder, G., Wisniewski, P. J., & Welch, G. F. (2021). Autonomous vehicle visual embodiment for pedestrian interactions in crossing scenarios: Virtual drivers in AVs for pedestrian crossing. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems,* Yokohama Japan, Article 304. https://doi.org/10.1145/3411763.3451626

Hager, J. C., & Ekman, P. (1979). Long-distance of transmission of facial affect signals. *Ethology and Sociobiology, 1*, 77–82. https://doi.org/10.1016/0162-3095(79)90007-4

Hamlet, C. C., Axelrod, S., & Kuerschner, S. (1984). Eye contact as an antecedent to compliant behavior. *Journal of Applied Behavior Analysis, 17*, 553–557. https://doi.org/10.1901/jaba.1984.17-553

Jaguar Land Rover. (2018). The virtual eyes have it. https://web.archive.org/web/20181203150349/https://www.jaguarlandrover.com/2018/virtual-eyes-have-it

Katz, A., Zaidel, D., & Elgrishi, A. (1975). An experimental study of driver and pedestrian interaction during the crossing conflict. *Human Factors*, *17*, 514–527. https://doi.org/10.1177/001872087501700510

Kong, X., Das, S., Zhang, Y., & Xiao, X. (2021). Lessons learned from pedestrian-driver communication and yielding patterns. *Transportation Research Part F: Traffic Psychology and Behaviour, 79,* 35–48. https://doi.org/10.1016/j.trf.2021.03.011

Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2016). *Joint attention in autonomous driving (JAAD).* arXiv. https://doi.org/10.48550/arXiv.1609.04741

Lampinen, J. M., Erickson, W. B., Moore, K. N., & Hittson, A. (2014). Effects of distance on face recognition: Implications for eyewitness identification. *Psychonomic Bulletin & Review, 21*, 1489–1494. https://doi.org/10.3758/s13423-014-0641-2

Lee, Y. M., Madigan, R., Giles, O., Garach-Morcillo, L., Markkula, G., Fox, C., Camara, F., Rothmueller, M., Vendelbo-Larsen, S. A., Holm Rasmussen, P., Dietrich, A., Nathanael, D., Portouli, V., Schieben, A., & Merat, N. (2021). Road users rarely use explicit communication when interacting in today's traffic: Implications for automated vehicles. *Cognition, Technology & Work, 23*, 367–380. https://doi.org/10.1007/s10111-020-00635-y

Löcken, A., Golling, C., & Riener, A. (2019). How should automated vehicles interact with pedestrians? A comparative analysis of interaction concepts in virtual reality. *Proceedings of the 11th International Conference Automotive User Interfaces*, Utrecht, the Netherlands, 262–274. https://doi.org/10.1145/3342197.3344544

Mahadevan, K., Somanath, S., & Sharlin, E. (2018). Communicating awareness and intent in autonomous vehicle-pedestrian interaction. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal, Canada, Article 429. https://doi.org/10.1145/3173574.3174003

Malmsten Lundgren, V., Habibovic, A., Andersson, J., Lagström, T., Nilsson, M., Sirkka, A., Fagerlönn, J., Fredriksson, R., Edgren, C., Krupenia, S., & Saluäär, D. (2017). Will there be new communication needs when introducing automated vehicles to the urban context? In N. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation. Advances in intelligent systems and computing* (pp. 485–497). Springer. https://doi.org/10.1007/978-3-319-41682-3_41

Martin, W. W., & Jones, R. F. (1982). The accuracy of eye-gaze judgement: A signal detection approach. *British Journal of Social Psychology, 21*, 293–299. https://doi.org/10.1111/j.2044-8309.1982.tb00551.x

Merat, N., Louw, T., Madigan, R., Wilbrink, M., & Schieben, A. (2018). What externally presented information do VRUs require when interacting with fully Automated Road Transport Systems in shared space? *Accident Analysis & Prevention, 118*, 244–252. https://doi.org/10.1016/j.aap.2018.03.018

Moore, D., Currano, R., Strack, G. E., & Sirkin, D. (2019). The case for implicit external human-machine interfaces for autonomous vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Utrecht, the Netherlands, 295–307. https://doi.org/10.1145/3342197.3345320

Morgan, C. J., Lockard, J. S., Fahrenbruch, C. E., & Smith, J. L. (1975). Hitchhiking: Social signals at a distance. *Bulletin of the Psychonomic Society*, *5*, 459–461. https://doi.org/10.3758/BF03333299

Myllyneva, A., & Hietanen, J. K. (2016). The dual nature of eye contact: To see and to be seen. *Social Cognitive and Affective Neuroscience, 11*, 1089–1095. https://doi.org/10.1093/scan/nsv075

Nathanael, D., Portouli, E., Papakostopoulos, V., Gkikas, K., & Amditis, A. (2019). Naturalistic observation of interactions between car drivers and pedestrians in high density urban settings. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, and Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International*

Ergonomics Association (pp. 389–397). Springer. https://doi.org/10.1007/978-3-319-96074-6_42

National Highway Traffic Safety Administration. (2020). *Traffic safety facts: 2018 data. Pedestrians* (DOT HS 812 850). https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812850

Norman, D. A. (1992). *Turn signals are the facial expressions of automobiles.* Addison-Wesley Publishing Company.

Núñez Velasco, J. P., Lee, Y. M., Uttley, J., Solernou, A., Farah, H., Van Arem, B., Hagenzieker, M., & Merat, N. (2021). Will pedestrians cross the road before an automated vehicle? The effect of drivers' attentiveness and presence on pedestrians' road crossing behavior. *Transportation Research Interdisciplinary Perspectives, 12*, Article 100466. https://doi.org/10.1016/j.trip.2021.100466

Onkhar, V., Bazilinskyy, P., Stapel, J. C. J., Dodou, D., Gavrila, D., & De Winter, J. C. F. (2021). Towards the detection of driver–pedestrian eye contact. *Pervasive and Mobile Computing, 76*, Article 101455. https://doi.org/10.1016/j.pmcj.2021.101455

Oudshoorn, M., De Winter, J. C. F., Bazilinskyy, P., & Dodou, D. (2021). Bio-inspired intent communication for automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, 80, 127–140. https://doi.org/10.1016/j.trf.2021.03.021

Pennycooke, N. (2012). *AEVITA: Designing biomimetic vehicle-to-pedestrian communication protocols for autonomously operating & parking on-road electric vehicles* [Doctoral dissertation, Massachusetts: Institute of Technology]. https://dspace.mit.edu/handle/1721.1/77810

Ranasinghe, C., Holländer, K., Currano, R., Sirkin, D., Moore, D., Schneegass, S., & Ju, W. (2020). Autonomous vehicle-pedestrian interaction across cultures: Towards designing better external Human Machine Interfaces (eHMIs). *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, HI. https://doi.org/10.1145/3334480.3382957

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Ren, Z., Jiang, X., & Wang, W. (2016). Analysis of the influence of pedestrians' eye contact on drivers' comfort boundary during the crossing conflict. *Procedia Engineering*, *137*, 399–406. https://doi.org/10.1016/j.proeng.2016.01.274

Rodríguez Palmeiro, A., Van der Kint, S., Vissers, L., Farah, H., De Winter, J. C. F., & Hagenzieker, M. (2018). Interaction between pedestrians and automated vehicles: A Wizard of Oz experiment. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 1005–1020. https://doi.org/10.1016/j.trf.2018.07.020

Rothenbücher, D., Li, J., Sirkin, D., Mok, B., & Ju, W. (2016). Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles. *Proceedings of the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication*, Columbia University, NY, 795–802. https://doi.org/10.1109/ROMAN.2016.7745210

Schneemann, F., & Gohl, I. (2016). Analyzing driver-pedestrian interaction at crosswalks: A contribution to autonomous driving in urban environments.

Proceedings of the 2016 IEEE Intelligent Vehicles Symposium, Gothenburg, Sweden, 38–43. https://doi.org/10.1109/IVS.2016.7535361

Smart. (2021). Technische gegevens: Je nieuwe smart in cijfers [Technical data: Your new smart in numbers]. https://www.smart.com/nl/nl/node/1119

Snyder, M., Grather, J., & Keller, K. (1974). Staring and compliance: A field experiment on hitchhiking. *Journal of Applied Social Psychology*, *4*, 165–170. https://doi.org/10.1111/j.1559-1816.1974.tb00666.x

Sripada, A., Bazilinskyy, P., & De Winter, J. C. F. (2021). Automated vehicles that communicate implicitly: Examining the use of lateral position within the lane. *Ergonomics*, *64*, 1416–1428. https://doi.org/10.1080/00140139.2021.1925353

Sucha, M., Dostal, D., & Risser, R. (2017). Pedestrian-driver communication and decision strategies at marked crossings. *Accident Analysis & Prevention*, *102*, 41–50. https://doi.org/10.1016/j.aap.2017.02.018

Sutton, T. M., Herbert, A. M., & Clark, D. Q. (2019). Valence, arousal, and dominance ratings for facial stimuli. *Quarterly Journal of Experimental Psychology, 72*, 2046–2055. https://doi.org/10.1177/1747021819829012

SWOV. (2020, July 27). Pedestrians [Fact sheet]. https://www.swov.nl/en/facts-figures/factsheet/pedestrians

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. *Journal of Human Evolution, 52*, 314–320. https://doi.org/10.1016/j.jhevol.2006.10.001

Uttley, J., Lee, Y. M., Madigan, R., & Merat, N. (2020). Road user interactions in a shared space setting: Priority and communication in a UK car park. *Transportation Research Part F: Traffic Psychology and Behaviour, 72*, 32–46. https://doi.org/10.1016/j.trf.2020.05.004

Vanderbilt, T. (2008). *Traffic. Why we drive the way we do (and what it says about us).* New York: Alfred A. Knopf.

Verma, H., Pythoud, G., Eden, G., Lalanne, D., & Evéquoz, F. (2019). Pedestrians and visual signs of intent: Towards expressive autonomous passenger shuttles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*, Article 107. https://doi.org/10.1145/3351265

Wang, P., Motamedi, S., Qi, S., Zhou, X., Zhang, T., & Chan, C. Y. (2021). Pedestrian interaction with automated vehicles at uncontrolled intersections. *Transportation Research Part F: Traffic Psychology and Behaviour, 77*, 10–25. https://doi.org/10.1016/j.trf.2020.12.005

World Health Organization. (2020, February 7). Road traffic injuries. https://web.archive.org/web/20200301154410/https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Yang, S. (2017). *Driver behavior impact on pedestrians' crossing experience in the conditionally autonomous driving context* [Doctoral dissertation, KTH Royal Institute of Technology]. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-220545

Chapter 3

Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions

This chapter has been published as:
Onkhar, V., Dodou, D., & De Winter, J. C. F. (2024). Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions. *Behavior Research Methods*, *56*, 4221–4238. https://doi.org/10.3758/s13428-023-02173-7

Abstract

Over the past few decades, there have been significant developments in eye-tracking technology, particularly in the domain of mobile, head-mounted devices. Nevertheless, questions remain regarding the accuracy of these eye-trackers during static and dynamic tasks. In light of this, we evaluated the performance of two widely used devices: Tobii Pro Glasses 2 and Tobii Pro Glasses 3. A total of 36 participants engaged in tasks under three dynamicity conditions. In the “seated with a chinrest” trial, only the eyes could be moved; in the “seated without a chinrest” trial, both the head and the eyes were free to move; and during the walking trial, participants walked along a straight path. During the seated trials, participants’ gaze was directed towards dots on a wall by means of audio instructions, whereas in the walking trial, participants maintained their gaze on a bullseye while walking towards it. Eye-tracker accuracy was determined using computer vision techniques to identify the target within the scene camera image. The findings showed that Tobii 3 outperformed Tobii 2 in terms of accuracy during the walking trials. Moreover, the results suggest that employing a chinrest in the case of head-mounted eye-trackers is counterproductive, as it necessitates larger eye eccentricities for target fixation, thereby compromising accuracy compared to not using a chinrest, which allows for head movement. Lastly, it was found that participants who reported higher workload demonstrated poorer eye-tracking accuracy. The current findings may be useful in the design of experiments that involve head-mounted eye-trackers.

3.1. Introduction

Eye-tracking, though seemingly a modern technique, is in fact by no means new, having been around in various forms for over 100 years (Płużyczka, 2018). As with any technology, its design and performance have improved over the decades, from invasive rods placed on users’ corneas and connected to sound-producing drums (Lamare, 1892) to noninvasive, remote screen-based systems as well as lightweight glasses equipped with infrared cameras. These contemporary eye-tracking devices are used in a variety of research fields, including psychology, marketing, art, sports, and human-computer interaction to investigate visual attention, cognitive processes, and user experience (Kredel et al., 2017; Meißner et al., 2019; Rosenberg & Klein, 2015).

However, with regard to the accuracy of eye-trackers, a mismatch has repeatedly been noted between their observed values and those reported in manufacturer specifications (Ehinger et al., 2019; Holmqvist, 2017; Morgante et al., 2012; Stuart et al., 2016). Therefore, there is a need to determine the accuracy of eye-trackers used in human subject research.

A variety of previous studies have evaluated the accuracy of eye-tracking technology. For example, Serchi et al. (2014) evaluated the Tobii TX300 remote (i.e., screen-based) eye-tracker. Their experiment involved four participants looking at a white dot that appeared sequentially for 2 seconds each in a grid of 13 dots, while

standing at different distances, or walking on a treadmill at a speed of 0.6 m/s or 1.1 m/s. They reported that the distance between the participant and the eye-tracker cameras was a critical factor in determining accuracy, but whether the participant was walking or not had little influence.

Recognizing the need to benchmark affordable remote eye-tracker models, Gibaldi et al. (2017) evaluated the Tobii EyeX, a low-cost eye-tracker attached to a screen. In their experiment, 15 participants were seated 0.7 m from the screen with a chinrest, and looked at circular targets appearing for 2 seconds on a grid, in random order. Their results showed that accuracy decreased with target eccentricity.

In a large-scale effort, Holmqvist (2017) assessed 12 eye-trackers with up to 194 participants per eye-tracker. They also analyzed participant characteristics that could potentially impact data quality, including eye color, eye makeup, pupil size, screen position, and the use of glasses. An unexpected result was that, compared to earlier studies of the same research group (Lång et al., 2011; Nyström et al., 2013), accuracy was worse. Several possible explanations were provided, including the low luminance of the environment, inexperienced or unmotivated experimenters, and the wide variety of participants. Generally, accuracy was found to be poorer for larger target eccentricities, and for certain participant characteristics, such as blue eyes (in infrared, a blue iris appears darker than a brown iris), the use of makeup (mascara can be mistaken by the eye-tracker for a pupil because they both appear dark), glasses with anti-reflective coating, scratches, or dirt, and soft contact lenses (which may generate infrared reflections).

Concerning mobile, also known as head-mounted or wearable, eye-trackers, Stuart et al. (2016) assessed the accuracy of a Dikablis model, developed by Ergoneers. Thirty-four older participants (14 with Parkinson's disease, 20 without) gazed at two targets placed 5°, 10°, and 15° apart in time with a metronome of 1 Hz for 20 seconds while seated and using a chinrest, standing and not moving their head, or walking on a treadmill. Accuracy was defined as the bias of saccade amplitude, with bias, in turn, defined as the difference between known target distance, i.e., eccentricity, and median saccade amplitude. It was evident that accuracy was poor and depended on the target eccentricity, but it did not seem to be significantly affected by whether participants sat, stood, or walked. The authors noted that the accuracy observed in the study was considerably worse than the 0.5° accuracy claimed by the manufacturer. They also observed that accuracy was better among participants with no visual correction as compared to those with correction.

Niehorster et al. (2020) conducted an experiment to investigate how accurately four wearable eye-trackers (Tobii Pro Glasses 2, SMI Eye Tracking Glasses 2.0, Pupil Labs Pupil in 3D mode, and Pupil Labs Pupil with Grip gaze estimation algorithm) recorded gaze when the glasses slipped on participants' noses. Nine participants looked at (the center of) a grid containing eight ArUco markers at a distance of 1.5

m, while pronouncing vowels, making facial expressions, or moving the eye-trackers on their face using their hands. The authors observed that while the gaze estimates of the Tobii and Grip remained stable, the other two systems exhibited significant increases in gaze deviation when performing such movements, which raises concerns that they may not be suitable for use in dynamic scenarios.

Pastel et al. (2021) assessed the accuracy of the Eye Tracking Glasses 2.0 (SMI, Germany). Twenty-one participants were seated in front of a computer screen, used a chinrest, and sequentially performed three tasks: looking at stationary targets appearing at four locations, tracking a target moving in the shape of an infinity loop, and looking straight ahead at stationary targets at different distances. In line with previous studies, accuracy was found to be poorer for more eccentric gaze targets.

Finally, Hooge et al. (2022) compared six different eye-trackers (Pupil Core 3D, Pupil Invisible, SMI Eye Tracking Glasses 2 60 Hz, SeeTrue, Tobii Pro Glasses 2, Tobii Pro Glasses 3) in various conditions (e.g., standing still, walking along a circle, jumping). The results of four participants showed that the best accuracy occurred for the standing-still condition, but substantially poorer accuracy was obtained for walking, skipping, and jumping.

To summarize, a number of studies on remote and mobile eye-trackers have shown that eye-trackers are less accurate for targets at greater eccentricities (Gibaldi et al., 2017; MacInnes et al., 2018; Niehorster et al., 2020; Pastel et al., 2021; Stuart et al., 2016). There is less consensus on the effect of dynamic tasks, with earlier research (Serchi et al., 2014; Stuart et al., 2016) reporting no large differences between sitting, standing, and walking, while the recent study by Hooge et al. (2022) showed a clear reduction in accuracy with increased dynamicity from standing still to walking, skipping, and jumping. It should be noted, however, that Serchi et al. (2014) used a remote eye-tracker, which is normally not used while walking. Another factor to take into consideration in assessing the accuracy of mobile eye-trackers concerns the automated localization of the visual target in the camera image. This has been done by mapping the camera image to a reference image with the help of feature matching (MacInnes et al., 2018) or ArUco markers (Ehinger et al., 2019; Niehorster et al., 2020), or alternatively, by identifying the colored fixation target in the camera image (Hooge et al., 2022). These methods may introduce errors, depending on the method used. These challenges highlight the need for further research on the accuracy of mobile eye-trackers in dynamic tasks using appropriate computer-vision algorithms.

The current study investigates accuracy as a function of dynamicity by using two popular mobile eye-trackers: the Tobii Pro Glasses 2 and 3. The Tobii 2 is a widely used eye-tracker that has four eye cameras (2 per eye) and 12 illuminators (6 per eye), which are integrated into the frame of the glasses below and above the eyes. The Tobii 3 is a newer model with a more streamlined appearance resembling

conventional glasses. It uses four eye cameras (2 per eye) and 16 illuminators (8 per eye) that are integrated into the lenses instead, for better positioning and supposedly more robust eye-tracking.

In our study, we assessed the performance of Tobii Pro Glasses 2 and 3 under three distinct conditions: the first encompassing only eye movements, the second incorporating both head and eye movements, and the third involving a combination of body, head, and eye movements. Investigating these three conditions allows for a comprehensive understanding of the devices' performance under various realistic scenarios. Notably, our second condition represents an important (and until now, missing) bridge between seated, static trials and walking ones, two of the most commonly tested scenarios in research evaluating eye-trackers. We hypothesized that with each added layer of dynamicity, eye-tracking accuracy would worsen.

In addition to assessing eye-tracker accuracy for different task conditions, we evaluated how participant characteristics correlated with eye-tracker accuracy. We expected accuracy to be worse for participants who wore contact lenses and for participants with blue eyes because of their reduced contrast against pupils in infrared light (Holmqvist, 2017). Gender was not expected to be of influence (Holmqvist et al., 2022). Previous studies propose that eye-tracker accuracy might be affected by lighting conditions, considering that pupil diameter tends to vary in response to light (Hooge et al., 2021; Wyatt, 2010). According to documentation from Tobii, the accuracy of the Tobii 2 and Tobii 3 eye-tracking devices may be substantially compromised in environments with minimal lighting (1 lux) (Tobii AB, 2017b, 2022b). In the present study, although the lighting conditions were not as low, an investigation was conducted to determine the relationship between the recorded pupil diameter of the participants and the eye-tracker's accuracy. Finally, we used the NASA Task Load Index (TLX) questionnaire to understand the association between eye-tracker accuracy and facets of perceived workload (i.e., mental demand, physical demand, temporal demand, performance, effort, and frustration), thus making it possible to assess whether eye-tracker accuracy is purely software- and hardware-related or also tied to participant state.

3.2. Methods

3.2.1. Participants

Thirty-six participants (20 males, 16 females) between the ages of 21 and 38 years (mean: 27.19, *SD*: 3.07, median: 26 years) were recruited via social media and direct contact to take part in the experiment between December 13, 2021 and February 4, 2022. Most were PhD candidates (22 participants) or employees at the Delft University of Technology or elsewhere (8 participants). The remaining participants were a postdoctoral researcher (1 participant) and (former) students (5 participants). Only people with normal visual acuity, corrected-to-normal vision using contact lenses, or low refractive errors (such that prescription lenses were not required for daily life activities) were eligible to participate.

Precautions were also taken against the spread of COVID-19 (sanitization of participants' and experimenter's hands, surfaces, and equipment touched, and social distancing whenever possible). All individuals provided written informed consent. The research was approved by the Human Research Ethics Committee of the Delft University of Technology (reference number 1832).

3.2.2. Eye-trackers

Two head-mounted eye-trackers, the Tobii Pro Glasses 2 (firmware version 1.25.6-citronkola-0, head unit version 0.0.62) and the Tobii Pro Glasses 3 (firmware version 1.23.1+pumpa), were used to track participants' gaze at 50 Hz and 100 Hz, respectively, and a forward-facing scene camera in each recorded their field of view at 25 frames per second and a resolution of 1920×1080 pixels. Note that the Tobii 2 allows for 100 Hz recordings by alternating the measurements from each eye (Holmqvist et al., 2022; Niehorster et al., 2020), an approach not taken in the present study. The Tobii 2 was used without its detachable protective lenses.

3.2.3. Experiment Setup

The experiment was conducted indoors in a workplace setting, with the seated trials performed in a private office and the walking trials in a nearby corridor, both of which were illuminated by natural and overhead lighting.

3.2.3.1. Seated Trials

A pattern consisting of nine green dots was printed on white A1-size paper and attached on a wall; this arrangement included a central dot surrounded by eight equidistant peripheral dots, forming a circle with a diameter of 536 mm. Note that it was decided to use printed gaze targets instead of a digital display, such as a television screen, due to its portability and ease of setup, making it simpler to replicate the study.

A table of 1 m lateral width was placed against the wall, and a chinrest was clamped on the table's opposite side, at its longitudinal center. At this distance from the wall, each dot had an eccentricity of 15° from the center. The dots themselves had a visual span of approximately 1.1° (20 mm diameter). The selection of a 15° eccentricity was informed by considerations of user comfort and applicability to real-life tasks. In tasks involving target detection, humans typically employ eye movements for small target eccentricities, and incorporate head movements for larger eccentricities to reduce eye strain (Stahl, 1999). Stahl found a mean eyes-only range across participants to be 35.8° (median of 25.3°), which is consistent with our 30° range and the ranges used in tests by eye-tracker manufacturers (Tobii AB, 2017b, 2022b). These assumptions align with eye movements during naturalistic tasks, such as walking, where individuals tend to focus on targets by employing a combination of eye and head movements. Standard deviations of eyes-in-head angles typically range from 5° to 10°, depending on factors such as terrain

roughness (Bahill et al., 1975; Foulsham et al., 2011; Franchak et al., 2021; ’t Hart & Einhäuser, 2012).

The height of the chinrest was set prior to the experiment so that the central dot of the pattern was aligned with the experimenter’s eye level while seated in an office chair and using the chinrest. The height of the chinrest was not to be adjusted during the experiment, but participants were free to adjust the chair height to sit comfortably. The chinrest was unclamped and re-clamped to the table between trials, without compromising its preset height and position along the table. The chinrest was used without its removable forehead attachment, since it was not possible to press one’s forehead against it without having the eye-tracker collide with the setup. Two speakers were used in the seated trials to play audio instructions to guide participants’ gaze across the pattern. Figures 3.1 and 3.2 illustrate the seated trials with and without a chinrest, respectively.

Figure 3.1. Seated trial with chinrest.

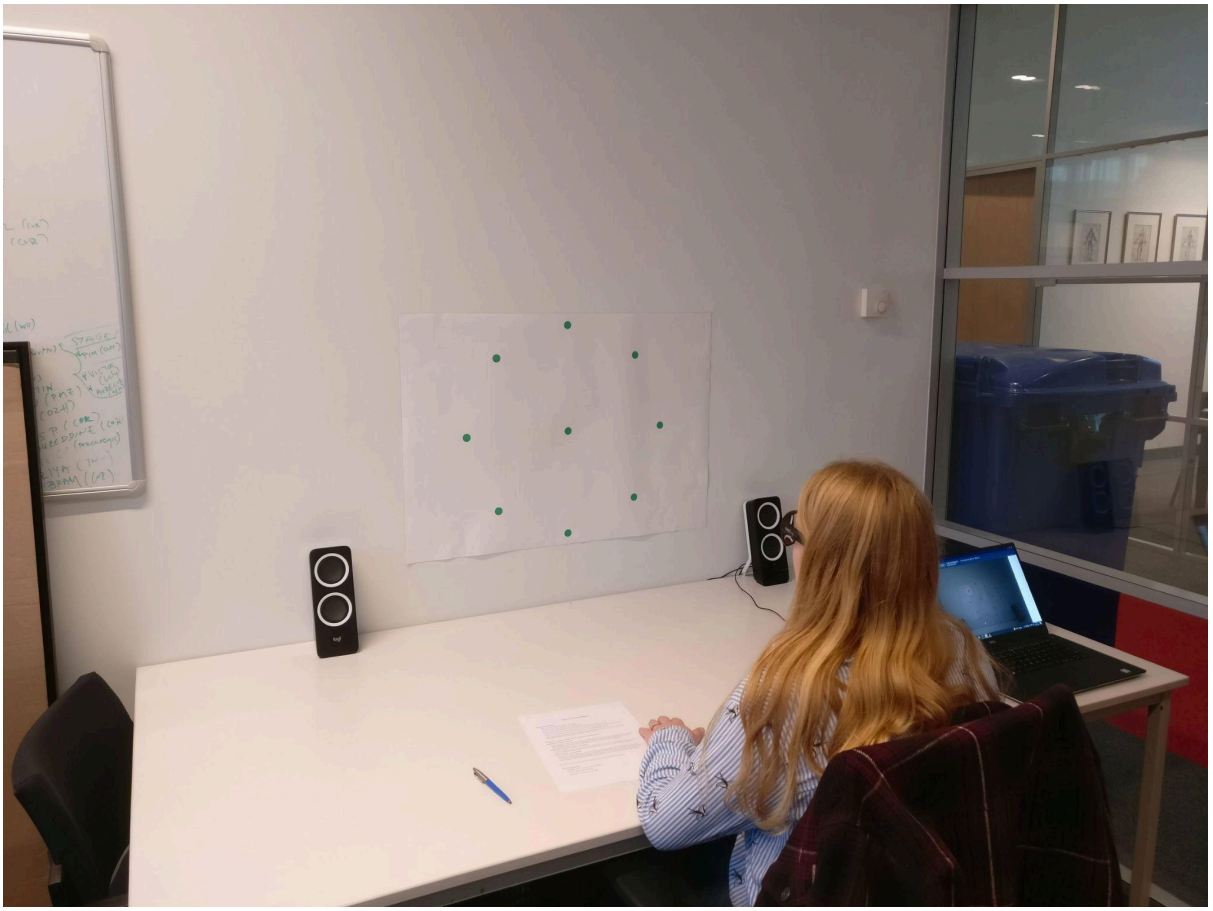

Figure 3.2. Seated trial without chinrest.

3.2.3.2. Walking Trials

A green bullseye (480 mm outer diameter) was printed on A1-size paper and mounted on a mobile whiteboard. The whiteboard was placed at one end of a corridor, on the edge of a 21.7-m long carpet. Participants would stand at the opposite end of the corridor before they commenced walking towards the bullseye during the trial. Figure 3.3 shows a walking trial in progress.

3.2.4. Experimental Design and Procedure

The experiment involved three types of trials for each of the two eye-trackers, resulting in a total of six trials as shown in Table 3.1. A blocked design was implemented, with each block using one eye-tracker. Half of the participants, specifically those with odd participant numbers, began with a block using the Tobii 2, followed by a block using the Tobii 3. The remaining participants with even participant numbers started with a block using the Tobii 3 and then moved on to a block using the Tobii 2. Within each block, the sequence of the three dynamicity conditions was random.

Figure 3.3. Walking trial.

Table 3.1
Experimental trials

Dynamicity condition	Eye-tracker	Permitted movements	Gaze target
Seated, with chinrest	Tobii 2	Eyes	Pattern of 9 green dots
Seated, without chinrest	Tobii 2	Eyes, head	Pattern of 9 green dots
Walking	Tobii 2	Eyes, head, body	Green bullseye
Seated, with chinrest	Tobii 3	Eyes	Pattern of 9 green dots
Seated, without chinrest	Tobii 3	Eyes, head	Pattern of 9 green dots
Walking	Tobii 3	Eyes, head, body	Green bullseye

Upon arrival, participants sanitized their hands, signed the consent form, and completed a questionnaire on their demographic data, eye color, and visual acuity. They were then briefed about the aim, procedure, and tasks of the experiment. Next, they put on one of the eye-trackers (depending on the predefined random order of trials assigned to them) and, if necessary, the eye-tracker nose pad size was adjusted for better comfort.

Participants' gaze was then calibrated using a bullseye card that the participant held at arm's length. A successful calibration was achieved when the participant's gaze marker sufficiently overlapped with the bullseye for a specified period of time, criteria

that were internally determined by the manufacturer's software. All participants achieved successful calibration and no participants were excluded because of failing to calibrate.

The robustness of the calibration was verified by asking the participant to move the card to multiple points of varying eccentricity (up, down, left, right), during which they looked at the card without rotating their head, using only their eyes. Recalibration was performed if there was insufficient overlap between the participant's gaze marker and the bullseye. After successful verification, participants were not permitted to adjust the eye-tracker's position on their faces until the upcoming trial was completed. Calibration and verification were performed before each trial, with participants standing in a designated area of the private office. Breaks were provided between trials if necessary.

In the seated trials with a chinrest, participants adjusted the chair height to sit comfortably, carefully placed their chin on the chinrest, and gazed at specific dots in the pattern (using only eye movements) for 12-second intervals each, following audio instructions in a synthesized female voice played in random order. The 12-second interval was chosen to ensure 10 seconds of available data per instructed dot (assuming it took participants no more than 2 seconds to respond to an instruction and focus on a new dot). The instructions were directions, each corresponding to a specific dot: "center", "top", "bottom", "left", "right", "top left", "top right", "bottom left", and "bottom right". The central dot was called out, and hence to be visited, three times (at the start, middle, and end of the trial), and the remaining dots were called out twice each (in random order), in a trial that lasted just under 4 minutes.

Similarly, in the seated trials without a chinrest, participants gazed at specific dots in the pattern for 12-second intervals each, as per the audio instructions. They were also instructed in advance by the experimenter to look at the dots as they might naturally do, i.e., they were made aware they had the freedom to rotate their head as well as their eyes. The chinrest was unclamped and placed aside beforehand. These trials also lasted approximately 4 minutes and involved three visits to the central dot (at the start, middle, and end of the trial) and two visits each to the remaining dots in random order.

In the walking trials, when the corridor was free of passers-by and disturbances, participants walked from one of its ends to the bullseye at the other end, while keeping their gaze fixed on the center of the target. They were asked to walk normally, which meant eye, head, and body movements were all permissible. When they had reached a close distance to the bullseye, participants turned around, walked back to their starting position, and repeated this exercise once more. These trials lasted approximately 1 minute.

After each block of trials, i.e., after completing the three types of trials with a specific eye-tracker, participants completed the NASA TLX questionnaire, which polled six facets of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. The responses were recorded on a horizontal scale with 21 ticks, with anchors at the ends representing *very low* and *very high*, respectively. For the performance item, the anchors used were *perfect* and *failure*, respectively. Participants then put on the other eye-tracker and repeated the entire procedure once more, at which point the experiment was finished, and they were free to leave. Before the arrival of the next participant, the experimenter disinfected all surfaces and equipment that came into physical contact using alcohol wipes.

3.2.5. Data Preprocessing

Once the experiment was completed by all participants, the raw eye-tracking data were exported as separate .xlsx files. In addition, .mp4 video files from the Tobii project folders were used. The analysis used the variables 'Gaze Point X' and 'Gaze Point Y', which represent the coordinates of the averaged gaze points for the left and right eyes in pixels, in the horizontal and vertical directions, respectively.

In the assessment of eye-tracker accuracy, it is important to focus on relevant accuracy indicators, rather than high-frequency jitter and blinks, which are commonly addressed in standard practice. Therefore, the data were filtered and blinks were removed. Specifically, the *x*- and *y*-data were passed through a moving median filter (e.g., De Winter et al., 2022; Jarodzka et al., 2012; Onkhar et al., 2021). The median filter had a 0.30-second interval and omitted missing data, i.e., any window containing missing values is the median of all non-missing elements in that window. A median filter removes noise and outliers while preserving edge information. That is, a median filter preserves fast movements like saccades, as opposed to smoothing filters, which would cause blurring of these rapid transitions.

Figure 3.4 illustrates the filtering applied. Finally, the mean *x*- and *y*-positions were computed per video frame (i.e., two measurements per frame for the Tobii 2 and four measurements per frame for the Tobii 3).

The target coordinates were automatically extracted from the recorded video frames of the scene camera. For the seated trials, an image filter was applied so that the green dots stood out more clearly from the background. Next, MATLAB's *imfindcircles* function (Yuen et al., 1990) was used to extract the nine dots (see Figures 3.1 and 3.2). Various heuristics with regard to the expected distances between the dots were applied, to ensure that the dots were appropriately labeled (e.g., "center", "top", "top right"). For the walking trials, lines were fitted to the edges of the red carpet (see Figure 3.3), and the intersection of the lines was used to estimate the approximate position of the bullseye. Next, the *imfindcircles* function was applied to estimate the coordinates of the bullseye in the scene camera image.

Finally, a median filter was applied to the estimated coordinates of the target, using a time interval of five video frames (0.20 s) to remove possible jitter.

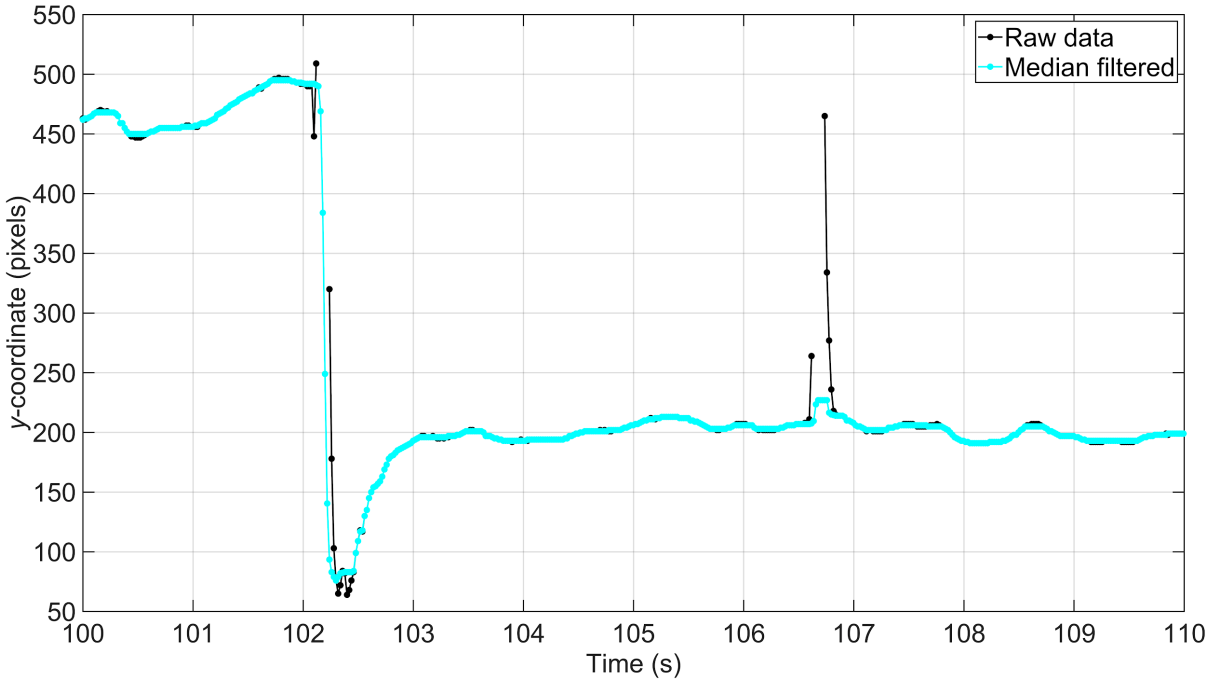

Figure 3.4. Filtering of the eye-tracking data for the Tobii 2. It can be seen that a blink at approximately 106.5 s was filled with data. Note that a *y*-coordinate of 1080 px corresponds to the bottom of the image (see also Figure 3.8).

Figure 3.5 illustrates the type of data collected in a seated trial with chinrest (top figure) and without chinrest (bottom figure), for a participant wearing the Tobii 2. The figure shows the continuously tracked coordinates of the nine dots, as well as the target dot (the audio instruction onsets were automatically extracted from the audio recorded by the eye-tracker), and the gaze *x*-coordinate. It can be seen that the participant tracked the target dot accurately, and in the condition without chinrest, also rotated their head, as indicated by the changes in the position of the dots (especially after being instructed regarding a new target dot).

Figure 3.6 shows the equivalent target and gaze data for a walking trial. In the walking trials, the bullseye was not steady in the scene camera image, but oscillated according to the gait of the participant.

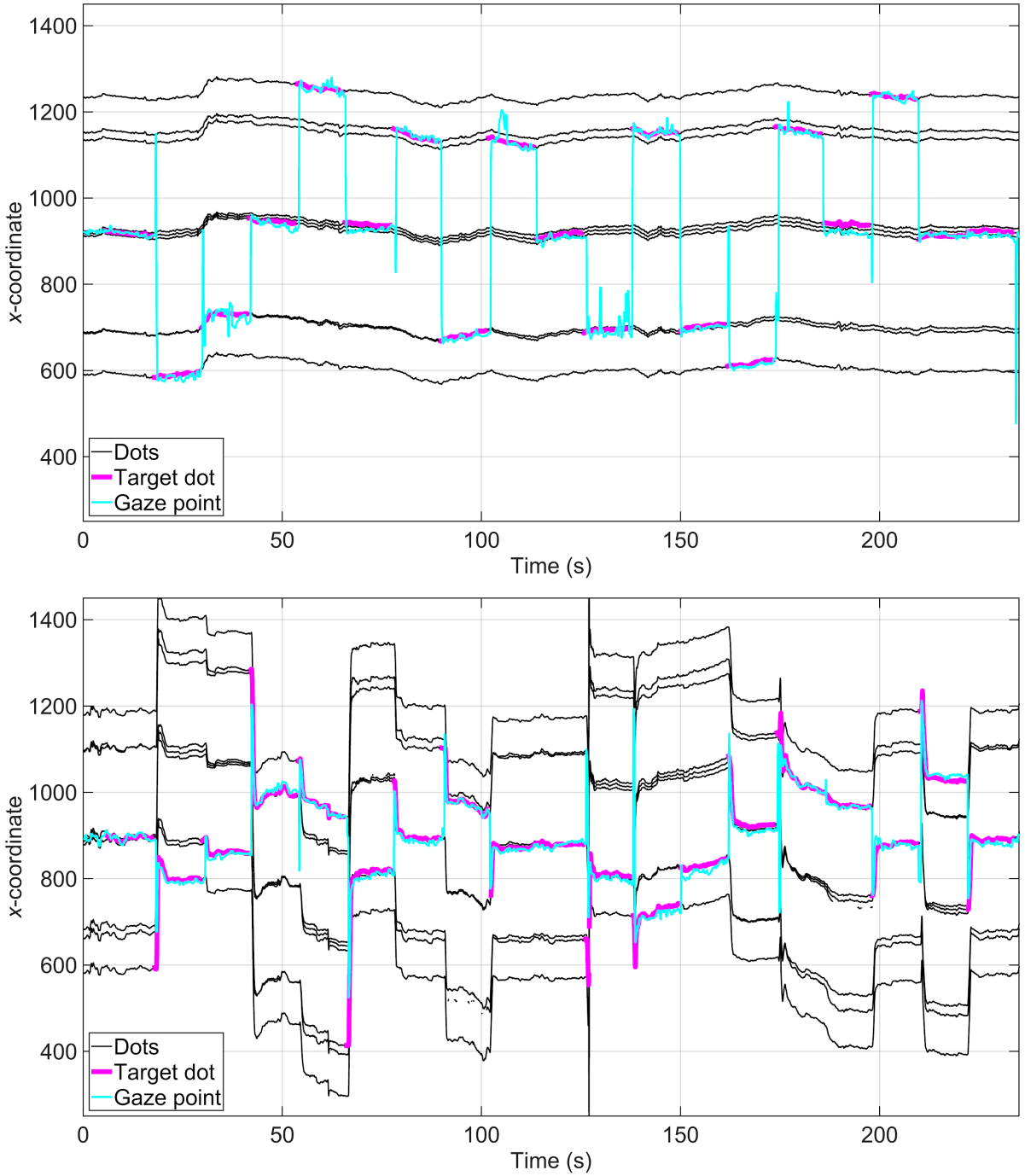

Figure 3.5. *x*-coordinate of the nine dots (in black) extracted using computer-vision, the instructed target dot highlighted in magenta, and a participant's gaze (in cyan), in the seated trials with the Tobii 2, with chinrest (top figure) and without chinrest (bottom figure).

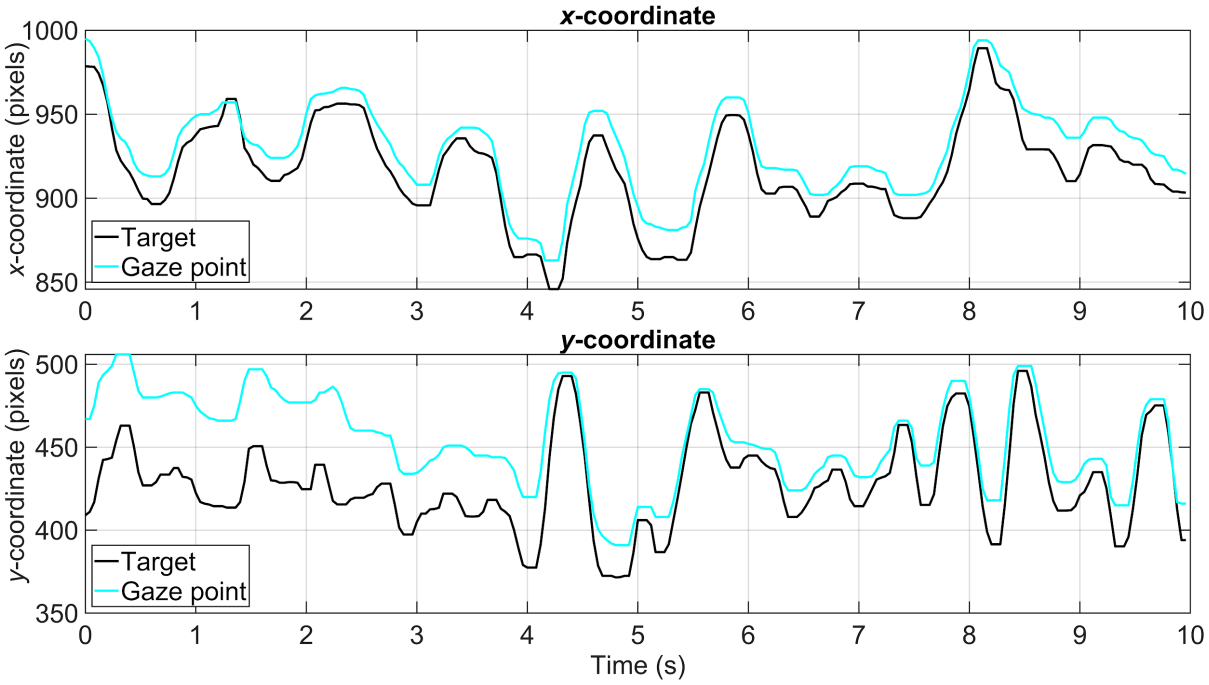

Figure 3.6. *x*- and *y*-coordinates of the estimated bullseye (in black) and filtered gaze (in cyan) for one participant in the walking trial (Tobii 3). The oscillating motion is caused by the participant's gait.

3.2.6. Computation of Angular Distance with Respect to Target

For each video frame, the angular distance between Tobii's instantaneous gaze point and the designated target was calculated. This was achieved by determining the angle between two vectors: the vector connecting the eyes to the gaze point, and the vector connecting the eyes to the target. The calculation was based on the dot product of the two vectors and the product of their magnitudes, as demonstrated in Eq. 1. The same approach has been previously employed in eye-tracking accuracy research (e.g., Aziz & Komogortsev, 2022, Eq. 4; Cercenelli et al., 2019, Eqs. 4 & 5; Mantiuk, 2017, Eq. 2; Xia et al., 2022, Eq. 26; and see Onkhar et al., 2021 using an equivalent formula using the cross-product and the dot-product).

$$\theta(i,j) = acos\left(\frac{(x_i-960)(x_j-960)+(y_i-540)(y_j-540)+VD^2}{\sqrt{(x_i-960)^2+(y_i-540)^2+VD^2}\cdot\sqrt{(x_j-960)^2+(y_j-540)^2+VD^2}}\right) \quad (1)$$

In Eq. 1, *x* and *y* are expressed in pixels, and (1,1) is the top-left corner of the image. *i* and *j* refer to the gaze coordinate and target coordinate at that moment, respectively. A constant is subtracted from the *x*- and *y*-coordinates to ensure that the angular distance is computed relative to the center of the scene camera image (e.g., Mantiuk, 2017). More specifically, for the *x*-coordinates, 960 pixels are subtracted, or half the screen width. For the *y*-coordinates, 540 pixels are subtracted, being half the screen height.

Viewing distance (*VD*) is a constant that relates to the magnification factor of the scene camera. If placing the Tobii closer (or farther) from the wall, the same translation in pixels corresponds to a proportionally smaller (or larger) translation in millimeters. *VD* can be interpreted as the virtual viewing distance, that is, the ratio between the distance between the Tobii scene camera and the wall in mm and the pixel size in mm. The *VD* parameter was determined by placing the Tobii approximately 1 m away from a wall with graph paper on it (Figures 3.7 & 3.8). The Tobii glasses were tilted so that the grid was vertically aligned with the borders of the camera image. Through manual and automated evaluations of the distances between grid lines, it was concluded that the camera view exhibited negligible distortion, aside from the outer few centimeters of the grid. Consequently, we opted to proceed without conducting a camera calibration designed to rectify such distortions. A screenshot of the camera view was made, and the distance in millimeters from the image center to various points (Figure 3.8) was determined with the help of the grid. Using the mean distances reported in Table 3.2 for a 400-pixel (px) eccentricity, *VD* was estimated to be 1132.4 px and 912.8 px for the Tobii 2 and 3, respectively. In other words, although both eye-trackers offered the same image resolution of 1920×1080 px, the Tobii 3 offered a larger field of view.

Figure 3.7. Setup for estimating the *VD* parameter. Graph paper consisting of 1×1 cm squares was stuck to the wall, and the Tobii was located at a distance of approximately 1 m from the wall.

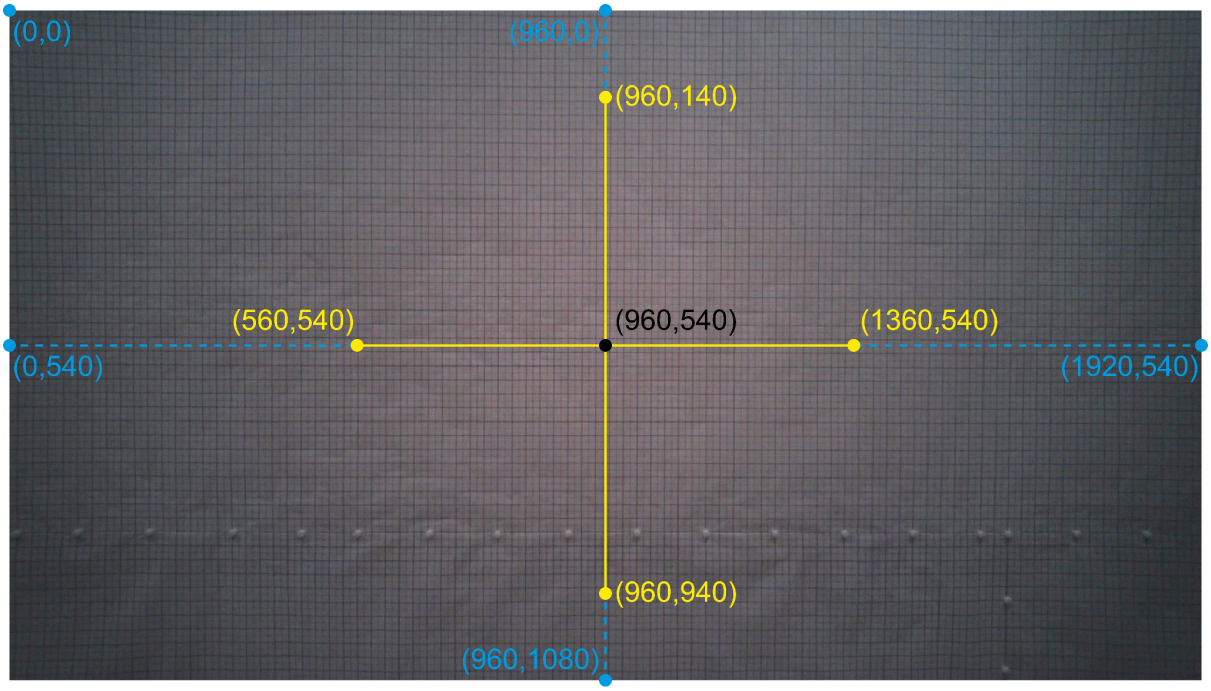

Figure 3.8. View of the scene camera pointing at the graph paper in the setup shown in Figure 3.7. The coordinates used in the analysis are shown. The distance to the edges (blue) was used to estimate the camera field of view, and the distance to points at 400 px eccentricity (yellow) was used to estimate the viewing distance (*VD*).

Table 3.2
Measured distances between coordinates shown in Figure 3.8 for the Tobii 2 and 3.

Model	Distance between scene camera and wall (mm)	Pitch angle of Tobii (°)	Distance between edge coordinates and center (°)	Distance between center and coordinates at 400 px eccentricity (°)
Tobii 2	1010	12	(0, 540): 846 mm (40.0°) (1920, 540): 850 mm (40.1°) (960, 0): 475 mm (25.2°) (960, 1080): 481 mm (25.5°)	(560, 540): 357 mm (19.5°) (1360, 540): 358 mm (19.5°) (960, 140): 354 mm (19.3°) (960, 940): 358 mm (19.5°)
Tobii 3	890	1	(0, 540): 950 mm (46.9°) (1920, 540): 970 mm (47.5°) (960, 0): 524 mm (30.5°) (960, 1080): 529 mm (30.7°)	(560, 540): 386 mm (23.4°) (1360, 540): 394 mm (23.9°) (960, 140): 390 mm (23.7°) (960, 940): 390 mm (23.7°)

Note. The distance between coordinates in mm was converted to degrees using *atan*(*d*/*D*), where *d* is the distance between the point and the image center in millimeters, and *D* is the distance between the scene camera and the wall in millimeters. For the Tobii 2, the horizontal and vertical FOVs were established to be 80.1° and 50.7°, which are close to the specifications reported by Tobii (82° and 52°; Tobii AB, 2017a). For the Tobii 3, the horizontal and vertical FOVs were estimated at 94.4° and 61.2°, again close to the manufacturer specifications (95° and 63°; Tobii AB, 2022a). The pitch angle was measured with a mobile app (RateFast, 2015).

In the seated trials, instructions to gaze at the dots were provided a total of 19 times (center dot: 3 times, other eight dots: 2 times each). The time interval between instructions to fixate on each dot was 12 seconds. Because the participant would

need time to shift focus from one dot to the next (see Figure 3.5), the first 2 seconds were discarded, and the angular distance was averaged over the remaining 10-second interval (i.e., 250 frames). For the walking trials, the participant walked up to the bullseye twice. Angular distance was averaged over a 10-second interval (see Figure 3.6) for each walk. The intervals for the walking trials were automatically extracted, with their end point being approximately when the red carpet went out of the scene camera's view and the starting point being 10 seconds before that moment.

The accuracy of the gaze of a participant in an experimental condition was calculated by computing the mean angular distance θ over the time intervals of that participant and condition. For the seated trials, 19 time intervals of 10 seconds each were available. For the walking trials, participants performed the task twice, and two time intervals of 10 seconds were used.

3.3. Results

Of the 36 participants, two participants (1 male, 1 female) completed only the Tobii 2 trials of the experiment due to a malfunction in the Tobii 3. Four other participants completed the experiment in two sessions spread across two separate days for the same reason. Furthermore, for one participant, the results for one condition (Tobii 3, without chinrest) were not available because of an error by the experimenter in carrying out the trials in their predefined order. Finally, for one of the participants in the walking trial with the Tobii 3, one of the two 10-second intervals was declared invalid due to an individual walking in front of the bullseye; consequently, the results of this trial rely on the data gathered from only one of the two time-intervals.

3.3.1. Accuracy of the Eye-Trackers

First, we determined the accuracy of the Tobii 2 and 3, where accuracy refers to the mean angular distance from the target. Table 3.3 shows the accuracy, averaged across participants and all nine dots, for the six experimental conditions. According to a two-way repeated-measures analysis of variance (ANOVA), there was a significant effect of Tobii model, $F(1,32) = 31.7$, $p < 0.001$, partial $\eta^2 = 0.50$, and of the level of dynamicity (i.e., with chinrest, without chinrest, or walking), $F(2,64) = 5.25$, $p = 0.008$, partial $\eta^2 = 0.14$. There was also a significant Tobii model × dynamicity interaction, $F(2,64) = 5.25$, $p = 0.008$, partial $\eta^2 = 0.14$.

Post-hoc paired-samples *t*-tests showed that for the Tobii 2, the chinrest condition yielded significantly poorer accuracy than the without-chinrest condition ($p = 0.001$) but not compared to the walking condition ($p = 0.123$). Furthermore, the without-chinrest condition yielded significantly better accuracy than the walking condition ($p = 0.003$). On the other hand, for the Tobii 3, there was no significant difference between the chinrest condition and the without-chinrest condition ($p = 0.134$) or the walking condition ($p = 0.859$). Also, the without-chinrest condition yielded no significant difference from the walking condition ($p = 0.404$). Upon

comparison of the two eye-trackers, it was observed that the Tobii 3 had significantly better accuracy than the Tobii 2 for the chinrest condition ($p < 0.001$) and walking condition ($p < 0.001$), but not for the without-chinrest condition ($p = 0.051$).

Table 3.3
Accuracy per experimental condition (in degrees) for all nine dots and for the central dot only. The mean, standard deviation (SD), and median across participants are reported.

		All dots			Only central dot			
Dynamicity condition	**Eye-tracker**	**Mean**	***SD***	**Median**	**Mean**	***SD***	**Median**	***n***
Seated, with chinrest	Tobii 2	2.77	1.49	2.40	1.44	1.90	1.04	36
Seated, without chinrest	Tobii 2	1.99	1.45	1.55	1.58	1.96	1.03	36
Walking	Tobii 2	3.53	2.70	2.86				36
Seated, with chinrest	Tobii 3	1.78	1.09	1.54	1.21	0.79	0.96	34
Seated, without chinrest	Tobii 3	1.60	0.98	1.27	1.23	0.96	0.86	33
Walking	Tobii 3	1.74	0.90	1.59				34

Note. For the walking trials, there was only one dot (i.e., the bullseye).

Note that the accuracy of the eye-trackers in the seated trials, as presented above, was calculated by computing the average across all nine dots. Table 3.3 also shows the accuracy specifically for the center dot. It can be seen that the mean accuracy for the center dot is markedly better as compared to all dots. This difference in accuracy can also be seen in Figure 3.9, which shows the angular distance to each instructed dot in the seated trials. It can be observed that for both eye-trackers, the center dot was detected more accurately than the others. In particular, for the Tobii 2 with chinrest, the eccentric dots showed poor accuracy compared to the center dot.

3.3.2. Movement of the Gaze Target

Heatmaps were created to better understand the underlying causes of the relatively poor accuracy of the Tobii 2. The heatmaps, shown in Figure 3.10, were created by dividing the camera image into squares of 20×20 pixels. It can be seen that the target was on average higher in the scene camera for the Tobii 2 compared to the Tobii 3. It can also be seen that participants without a chinrest were inclined to center the target dot in their field of view. That is, for the chinrest condition, participants looked at the dots by turning only their eyes, in accordance with the instructions. In contrast, during the without-chinrest condition, they also turned their heads, reducing the need to turn their eyes towards large eccentricities. Similarly, for the walking trials, in which there was only one target, participants tended to keep the target in the center of their view.

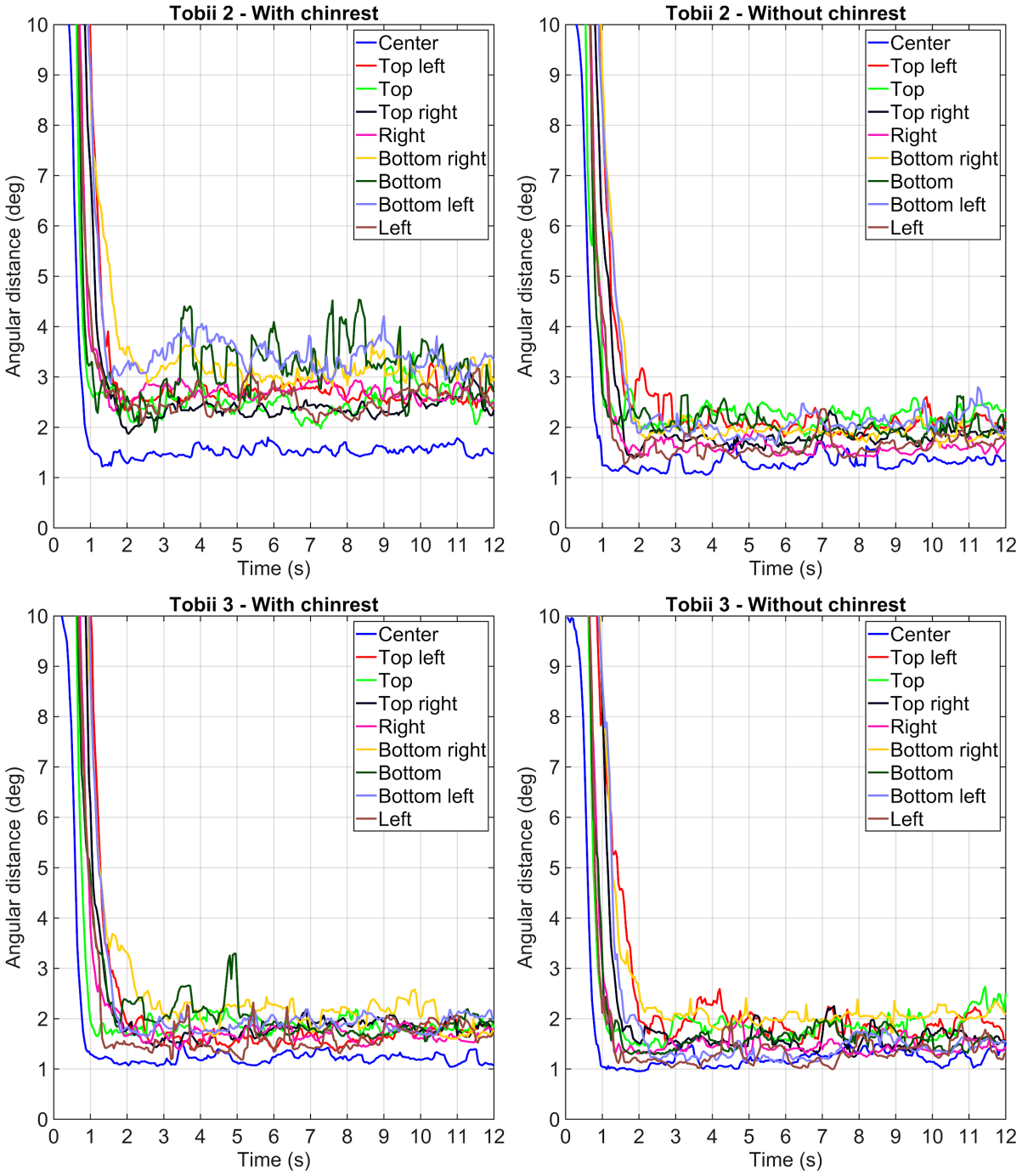

Figure 3.9. Angular distance to instructed dots, averaged across participants, for the Tobii 2 (top) and the Tobii 3 (bottom) and for the chinrest condition (left) and the without-chinrest condition (right). Note that the accuracy was determined for the last 10 seconds of the 12-second interval.

A restriction of movement was also associated with a larger amount of missing data. Specifically, the percentage of missing gaze data computed per video frame, and after filtering (see Methods) was 3.07, 1.88, and 0.12% for the chinrest, without-chinrest, and walking conditions of the Tobii 2, and 1.42, 0.58, and 0.12% for the Tobii 3, respectively. Before filtering, these values were 6.07, 3.96, and 1.41% for the Tobii 2, and 3.63, 2.43, and 1.38% for the Tobii 3.

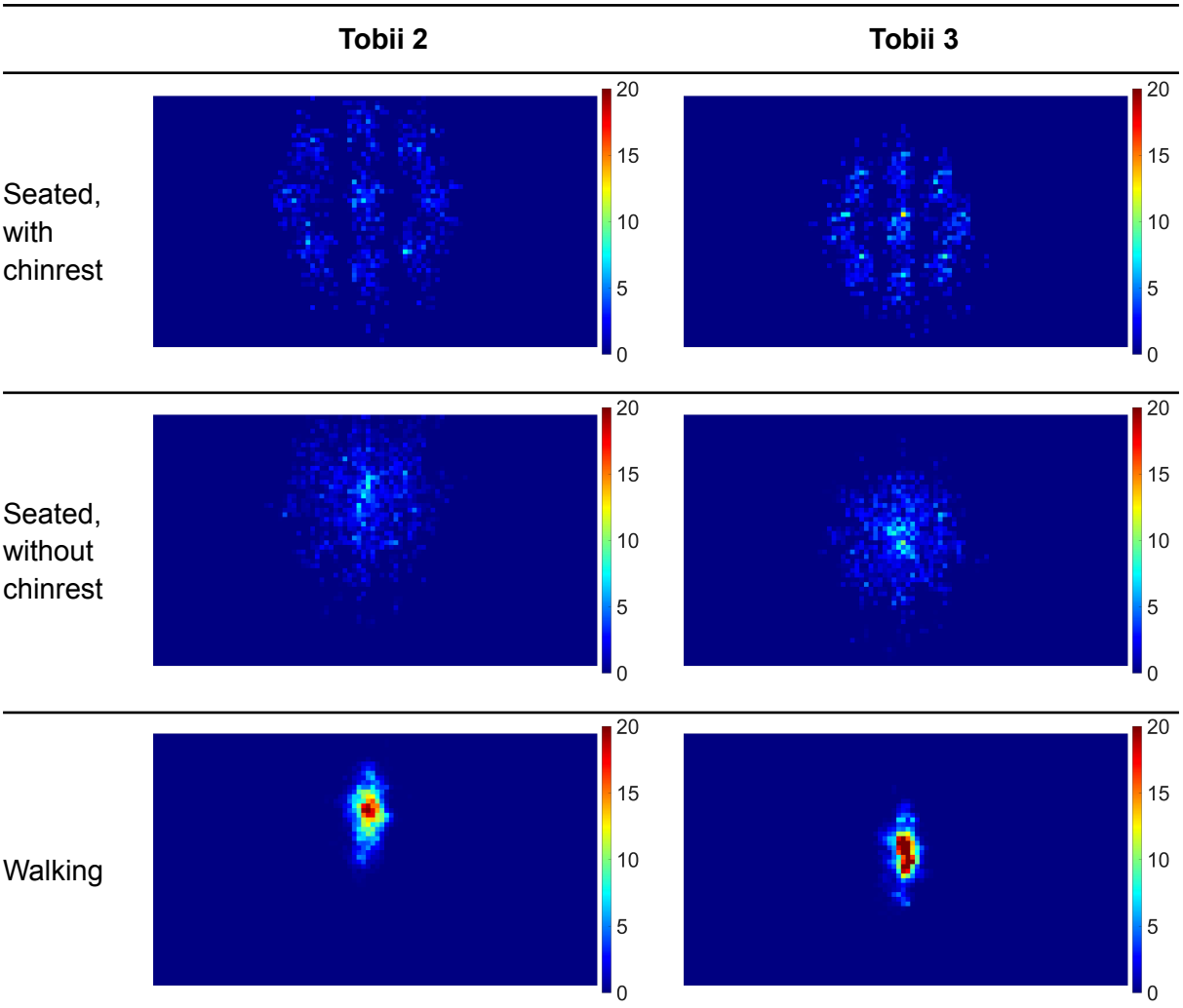

Figure 3.10. Heatmaps of the location of the targets (i.e., instructed dots for the seated trials, bullseye for the walking trials) across all 10-second intervals of all participants. The colormap represents the number of target points in the 20×20-pixel square, normalized so that the total of all squares equals 1000.

Subsequent to the above observations, an analysis of target speed within the camera image was conducted. Table 3.4 shows the mean speed of the identified target, averaged across the 10-second measurement interval. The instantaneous speed was computed using Eq. 1 for *x*- and *y*-coordinates for successive video frames. The results presented in Table 3.4 corroborate the efficacy of the chinrest in limiting head motion in comparison to the other conditions. There was no detectable effect of the Tobii model, which can be explained by the fact that the speed of the target is solely caused by participant movement, not by the eye-tracker itself. Specifically, according to a two-way repeated-measures ANOVA, there was no significant effect of the Tobii model, $F(1,32) = 0.09$, $p = 0.766$, partial $\eta^2 = 0.00$, but a strong effect of the level of dynamicity, $F(2,64) = 215.9$, $p < 0.001$, partial $\eta^2 = 0.87$. There was also no significant Tobii model × dynamicity interaction, $F(2,64) = 0.14$, $p = 0.873$, partial $\eta^2 = 0.00$.

Table 3.4
Mean speed of the target in the camera images per experimental condition (in degrees per second). The mean, standard deviation (SD), and median across participants are reported.

Dynamicity condition	Eye-tracker	Mean	*SD*	Median	*n*
Seated, with chinrest	Tobii 2	0.28	0.08	0.27	36
Seated, without chinrest	Tobii 2	0.83	0.70	0.68	36
Walking	Tobii 2	8.57	3.04	7.99	36
Seated, with chinrest	Tobii 3	0.28	0.08	0.26	34
Seated, without chinrest	Tobii 3	0.81	0.47	0.67	33
Walking	Tobii 3	8.77	3.88	8.39	34

3.3.3. Self-Reported Workload

To gain a deeper understanding of whether human workload experience is linked to eye-tracking accuracy, we examined the self-reported workload data from participants. Table 3.5 lists the perceived workloads when using the Tobii 2 and 3. It can be seen that the Tobii 3 resulted in statistically significantly lower physical demand, effort, and frustration compared to the Tobii 2.

Table 3.5
Mean and standard deviation of self-reported workload for Tobii 2 (n = 36) and Tobii 3 (n = 34). Also shown are the results of a paired-samples t-test.

	Tobii 2		Tobii 3		
	Mean	***SD***	**Mean**	***SD***	***t*-test**
TLX Mental demand (%)	29.7	22.8	27.9	19.9	$t(33) = 1.17$, $p = 0.249$
TLX Physical demand (%)	30.8	22.0	26.0	22.5	**$t(33) = 2.56$, $p = 0.015$**
TLX Temporal demand (%)	18.5	17.1	19.0	18.4	$t(33) = 0.19$, $p = 0.851$
TLX Performance (%)	30.3	21.6	29.1	21.8	$t(33) = 1.38$, $p = 0.176$
TLX Effort (%)	39.4	26.2	32.2	21.2	**$t(33) = 2.78$, $p = 0.009$**
TLX Frustration (%)	30.3	26.7	21.2	22.5	**$t(33) = 2.77$, $p = 0.009$**

Note. Scores were converted to a scale from 0% (minimum possible) to 100% (maximum possible). $p < 0.05$ is listed in boldface.

3.3.4. Individual Differences

In order to better understand the underlying factors that contribute to variations in eye-tracking accuracy, we examined individual differences among participants using the devices. The Pearson correlation between participants' overall accuracy (averaged across the dynamicity conditions) between the Tobii 2 and Tobii 3 was $r = 0.56$ ($p < 0.001$). This finding suggests that individual differences, such as unique eye characteristics or behavioral patterns, affected the performance of both eye-trackers; see Figure 3.11 for the tracking accuracy for the Tobii 2 and 3 per participant.

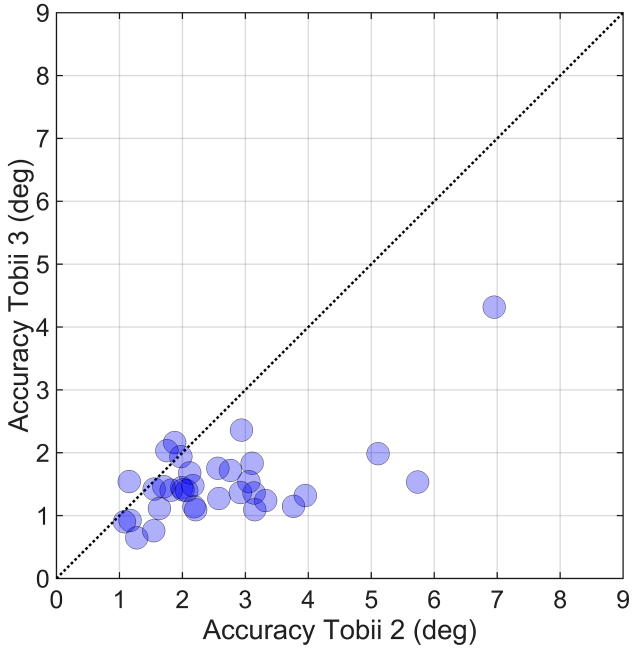

Figure 3.11. Overall eye-tracking accuracy (averaged across the three dynamicity conditions) for the Tobii 2 and 3. Each circular marker represents a participant (n = 34). The dotted line represents the line of equality.

Point-biserial correlations of participants' overall accuracy (z-transformed and subsequently averaged over all six conditions) did not reveal a statistically significant association with participant gender, whether or not the participant was wearing contact lenses, whether the participant had a particular eye color, and the participant's pupil diameter averaged across all conditions (see Table 3.6)[1]. However, there were significant associations with self-reported workload (physical demand, temporal demand, and frustration). These findings suggest that eye-tracking accuracy may be more influenced by participants' workload rather than their physical characteristics, highlighting the importance of considering human factors in eye-tracking research and technology development.

Finally, we investigated head movement as a possible factor influencing eye-tracking accuracy. Figure 3.12 illustrates that the speed of movement of the target was strongly person-specific, with some participants having a more dynamic gait than others regardless of Tobii model (r = 0.88, p < 0.001). However, the speed of the target did not correlate significantly with eye-tracking accuracy (r = -0.26, p = 0.133, n = 36 for the walking trials of the Tobii 2; r = -0.12, p = 0.509, n = 34 for the walking trials of the Tobii 3). That is, participants' eye-tracking accuracy was not significantly related to the speed of their head movement while walking.

[1] Our statement here applies to blue and brown eyes. However, we did find that the accuracy for hazel, green, gray, and amber eyes combined was statistically significantly better than for brown and blue eyes combined (see Table 3.6). The reasons for this are unclear but may involve confounding factors such as ethnic differences in eye shape. Because this effect is not clearly interpretable and only marginally significant (p = 0.040), we will not elaborate on it further.

Table 3.6

Means, sample sizes (n), and standard deviations of participants' characteristics and self-reported workload, together with the Pearson product-moment correlation coefficient with overall angular distance (n = 36).

	Mean	***n***	***r***	***p***
Gender (0: female / 1: male)	0.56	16 / 20	-0.09	0.619
Contact lenses (0: no / 1: yes)	0.22	28 / 8	0.16	0.341
Brown eyes (0: no / 1: yes)	0.39	22 / 14	0.19	0.276
Blue eyes (0: no / 1: yes)	0.14	31 / 5	0.23	0.172
Other eye color (0: no / 1: yes)	0.47	19 / 17	-0.34	**0.040**
	Mean	***SD***	***r***	***p***
Pupil diameter (mm)	3.80	0.66	0.06	0.732
TLX Mental demand (%)	28.5	20.4	-0.13	0.456
TLX Physical demand (%)	28.3	21.1	0.41	**0.013**
TLX Temporal demand (%)	18.3	17.1	0.46	**0.005**
TLX Performance (%)	29.2	21.1	0.09	0.598
TLX Effort (%)	35.8	22.3	0.32	0.056
TLX Frustration (%)	25.9	22.8	0.34	**0.040**

Note. Scores for the NASA TLX were averaged across the two eye-trackers (only the Tobii 2 in two participants), and converted to a scale from 0% (minimum possible) to 100% (maximum possible). $p < 0.05$ is listed in boldface. For binary variables (gender, contact lenses, eye color), the Pearson product-moment correlation coefficient is equivalent to the point-biserial correlation coefficient. Other eye colors include hazel, gray, green, and amber.

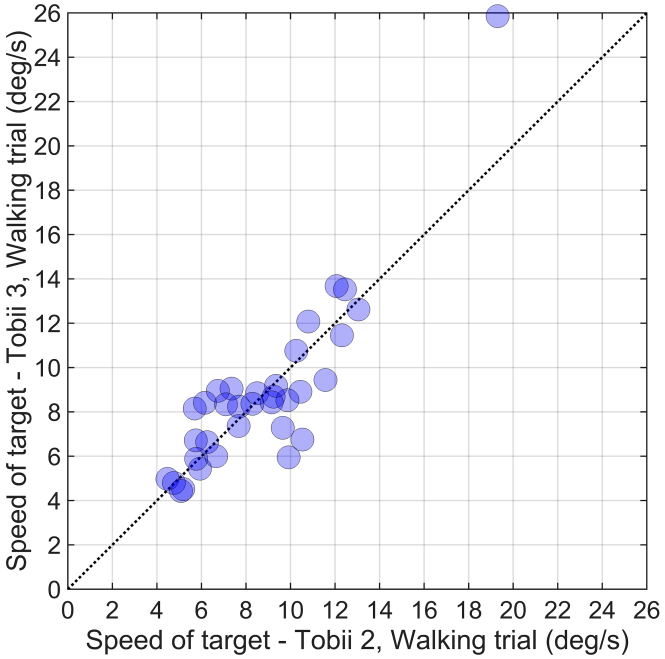

Figure 3.12. Mean speed of the target in the camera image in the walking trials for the Tobii 2 and the Tobii 3. Each circular marker represents a participant (n = 34). The dotted line represents the line of equality.

3.4. Discussion

In this study, the accuracy of Tobii Pro Glasses 2 and Tobii Pro Glasses 3 was compared under three distinct dynamicity conditions representing varying levels of freedom in movement: using a chinrest, without a chinrest, and while walking.

The accuracy of the Tobii 2 was found to be better without a chinrest, opposite to the standard approach for remote eye-trackers that recommends using a chinrest for enhanced accuracy (e.g., Holmqvist et al., 2011; Minakata & Beier, 2021; Niehorster et al., 2018). Interestingly, some researchers have used chinrests with mobile eye-trackers (Wang & Grossman, 2020; Werner et al., 2019). We initially hypothesized that eye-tracker accuracy would deteriorate with increasing levels of movement freedom; however, this did not turn out to be the case. The improved accuracy for the off-center dots without a chinrest can be ascribed to participants turning their heads, thereby reducing the need for gazing at large eccentricities. Eye-trackers generally exhibit better accuracy at smaller target eccentricities (MacInnes et al., 2018; Stuart et al., 2016), a notion supported by our seated trial results, wherein the central dot demonstrated the highest accuracy. Consequently, the use of a chinrest with a mobile eye-tracker is not recommended.

Our study also showed that the Tobii 3 exhibited statistically significant better accuracy than the Tobii 2 during the walking condition. This enhanced accuracy of the Tobii 3 may be attributed to the incorporation of extra illuminators and the more optimal placement of its eye-tracking cameras. Moreover, the Tobii 3's design possibly renders it more resistant to inaccuracies resulting from vibrations and other perturbations. Additionally, the Tobii 2 is characterized by a downward pitch angle of its scene camera (as also noted by Niehorster et al., 2020; Rogers et al., 2018; Thibeault et al., 2019, and which can be seen in the heatmaps in Figure 3.10). This deviation may compromise accuracy, as the central gaze point is situated at an eccentricity.

Another finding was that the difference in eye-tracking accuracy for the Tobii 3 during walking (averaging at 1.74°) and seated trials (averaging at 1.78° with chinrest and 1.60° without chinrest) was only small and not statistically significant. This result was unexpected, since research has suggested that eye-tracking accuracy is compromised during dynamic tasks such as walking (Hooge et al., 2022). In line with the above explanation for the chinrest vs. without-chinrest condition for the Tobii 2, a possible explanation is that participants were able to freely rotate their heads and bodies while walking, which helped to keep the target in the center of their field of view. This interpretation is supported by the heatmaps in Figure 3.10, which show relatively centralized gaze patterns for the walking trials. Additionally, the results suggest that participants were able to keep their gaze on the target while walking, regardless of how much head movement the participant exhibited. The effectiveness with which participants are able to track a target while walking can be attributed to the vestibular-ocular reflex, i.e., the reflexive eye movement that helps to stabilize

the gaze on a target while the head is moving (Dietrich & Wuehr, 2019; Moore et al., 1999). Future research could examine the factors that contribute to eye-tracking accuracy in dynamic tasks, including the role of head and body movements in maintaining a stable gaze. Additionally, it would be interesting to explore the generalizability of the findings to other mobile eye-trackers and different types of dynamic tasks, such as those requiring rapid head turns (e.g., crossing a road as a pedestrian, scanning a traffic scene as a driver) or tasks that require fast walking or running.

The present study examined the potential influence of several participant characteristics, namely gender, the use of contact lenses, eye color, and mean pupil diameter, on the overall accuracy of the eye-trackers. A sample of 36 participants was used, and the results did not reveal interpretable statistically significant associations with the accuracy of the eye-trackers. Previous research (Holmqvist, 2017; Nyström et al., 2013) suggested that contact lenses may reduce the accuracy of eye-trackers due to multiple corneal reflections. The results of our study did not support this hypothesis. It is possible that the Tobii eye-trackers were designed to accurately track eye movements even in the presence of these reflections. Future research could examine the effect of a larger variety of participant characteristics, including eye shape, and using larger sample sizes.

Our findings shown in Table 3.3 indicated that the mean accuracies across participants for the Tobii 2 and Tobii 3 in the chinrest condition were 2.77° and 1.78°, respectively. If selecting only the center target, the mean accuracies for the Tobii 2 and 3 are 1.44° and 1.21°, respectively. Our findings are in the ballpark of previous studies that have evaluated the Tobii 2, with mean accuracies between 1.19° and 5.25° for various target eccentricities (Niehorster et al., 2020). It is noteworthy, however, that even our accuracies for the center target alone are worse than those reported by Tobii, which reported a mean accuracy of 0.62° for the Tobii 2 in an unpublished test report (Tobii AB, 2017b), and 0.5° (for a central target) to 0.8° (for ‘common gaze angles’) in a Tobii 3 test report (Tobii AB, 2022b). There are several potential factors that could explain the discrepancies in accuracy observed between Tobii’s results and the current study:

- One factor is that Tobii AB (2017b) used another definition of accuracy. Following Holmqvist et al. (2011), we computed the angular distance per video frame, and subsequently calculated the mean angular distance, which is always a positive value, across those frames. Hence, in our case, accuracy represents the overall angular distance from the target. On the other hand, Tobii AB (2017b) and others (e.g., MacInnes et al., 2018) have defined accuracy as bias, a component of accuracy, which they calculated by determining the mean gaze point over a time interval and then computing the angular difference between this mean gaze point and the target. Bias, reflecting the average deviation from the true value, can present a misleadingly low value compared to the overall accuracy (absolute

error), by allowing overestimations and underestimations to offset each other. Therefore, bias should always be interpreted together with precision (which has been defined in different ways in the literature). For completeness, we offer bias and precision values for our experimental conditions in Appendices 3.A and 3.B. Bias values can be seen to be indeed smaller than the accuracy values presented in Table 3.3.

- Furthermore, in our seated trials, 3 out of 19 trials concerned the central dot, whereas the remaining 16 involved a dot at 15° eccentricity. In contrast, Tobii AB employed a more evenly distributed range of eccentricity values, which therefore constituted a less demanding evaluation. Specifically, the Tobii 2 test report (Tobii AB, 2017b) investigated one central dot, four dots situated at approximately 7° eccentricity, and another four dots at 10° eccentricity. Similarly, the Tobii 3 report (Tobii AB, 2022b) evaluated five dots within a 5° range, four dots at approximately 8° eccentricity, and four additional dots at 14° eccentricity. The fairness of each test may be a matter of debate: Despite both our study and Tobii's shared interest in targets at 15° eccentricity or below, our research was primarily concerned with targets at 15° eccentricity, the natural limit of eccentric gaze. In contrast, Tobii's evaluation included only a limited number of targets in close proximity to 15° eccentricity.
- Another consideration is the extended gaze duration employed in our seated trials, wherein accuracy was calculated over 10-second intervals, which is considerably longer than Tobii's 1-second window (Tobii AB, 2017b, 2022b). This decision was made to capture eye-tracker variability, although it introduced possible drawbacks, such as a challenge for participants to maintain focus. Figure 3.9 shows no evident difficulty in sustaining attention, as the mean angular distance across participants remained approximately constant throughout the measurement interval. Furthermore, it is important to note that large angular deviations should not necessarily be ascribed to participant inattention; eye-tracker inaccuracies themselves may also be a contributing factor.
- Another potential explanation for the better accuracy reported by Tobii AB (2017b) is that they may have included participants whose eyes were better trackable. Indeed, it should be noted that the median value of participants, as presented in Table 3.3, is lower than the mean value. This observation suggests that a small number of participants may be responsible for a disproportionately large portion of the observed inaccuracy. This point was also highlighted by Holmqvist (2017), who noted: "*we did our best to record all sorts of participants with troublesome features, while a typical study would try to exclude participants with glasses, mascara, squinting, unsuitable pupil sizes and other issues already during recruitment*" (p. 20).
- Also, based on correspondence received from Tobii AB in response to a preprint of this work, it was proposed that differences in firmware might have partially accounted for the observed discrepancy. In Tobii's tests of the Tobii 2, firmware 1.16.1 was used (Tobii AB, 2017b), whereas our study used the latest version accessible to users, 1.25.6-citronkola-0. Likewise, for the Tobii 3, our experiment

used the most recent firmware version available at the time, 1.23.1+pumpa, while Tobii's report, published later that year, made use of 1.28.1-granskott (Tobii AB, 2022b).

- As a final point, the methods of filtering and data processing implemented could potentially have had an influence. We developed the filtering algorithm to adequately handle artifacts such as blinks (see Figure 3.4), while also providing a robust estimate in the event of extensive missing data and jitter. In this sense, it should be noted that the filtering improves the apparent accuracy of the eye-tracker by an average of about 0.14° compared to using unfiltered data. In addition, it should be noted that during the seated trials with the Tobii 2, the top target occasionally disappeared from the camera view due to the previously mentioned tilt angle of the eye-tracker. This resulted in a gap in the measurement data, which in this paper is not considered as missing eye-tracking data or inaccurate measurement. The angular deviation also occasionally assumed very high values, which has a relatively large influence on the accuracy. One possible way to address this would be to define accuracy not as the mean angular deviation, as we have done, but as the median angular deviation. When this is done, the accuracy improves by approximately 19% compared to the values in Table 3.3.

Our study found that Tobii 3 had more accurate eye-tracking abilities, and also yielded statistically significantly lower physical demand, effort, and frustration than the Tobii 2, something which may be due to its more ergonomic design. In support of this, the experimenter noted that some participants reported partial obstruction of their view by the corners of the Tobii 2 frame when settling into the chinrest ahead of those trials and gazing eccentrically at the "left" and "right" target dots. An accuracy-workload correlation was also found at the level of participants, with participants who had better eye-tracking accuracy experiencing lower physical demand, temporal demand, and frustration. A possible explanation for the latter correlations is that participants who were less motivated and more fatigued exhibited increased bodily movement and a decreased ability to keep their eyes on the target. It is also possible that participants who had experienced experimental nuisances, such as calibration failures, experienced higher temporal demand and frustration. A recommendation that may follow from the above workload-accuracy correlations is that researchers who use eye-tracking may wish to consider measures to reduce workload in order to improve eye-tracking accuracy. For example, researchers should use concise instructions and provide adequate breaks.

Eye-tracker calibration verification in our study left some room for improvement. Although the calibration was rigorous, the experimenter was still responsible for subjectively verifying the accuracy of the calibration. It should also be noted that the trials in our experiment did not feature major disturbances such as wind, sunlight, or vibrations. In trials that may involve repeated facial movements or disturbances, or repositioning of the eye-tracker glasses, accuracy is likely to worsen. That said,

compared to other eye-trackers, the Tobii glasses have been found to be relatively resistant to accuracy degradation caused by slippage (Niehorster et al., 2020).

It is worth noting that in our experiment, the chinrest was used without a forehead attachment, as the participants' faces, when wearing the eye-tracker, could not be accommodated within the setup otherwise. Evidence from the scene camera indicated that participants' head movements were minimal during the chinrest trials (see Figure 3.5 for an example of one trial, which demonstrates limited movement of the dots, and Table 3.4 for numerical results). Consequently, it can be reasonably inferred that the findings of this study would not deviate significantly if the participants' heads were fully restrained.

Another limitation of the current study is that it only examined angular distances from a target in a number of standard tasks. Future research could examine the capabilities of mobile eye-trackers for a more comprehensive set of measures (see Ehinger et al., 2019, who developed a test battery consisting of 10 tasks to evaluate the Pupil Labs glasses eye-tracker).

3.5. Conclusions

This study showed that the Tobii Pro Glasses 3 yields better eye-tracking accuracy than the Tobii Pro Glasses 2 during walking. Furthermore, for the Tobii 2, not restraining the head yielded better eye-tracking accuracy than when a chinrest was used. Finally, participants who experienced higher workload exhibited poorer eye-tracking accuracy, which suggests that the observed eye-tracking accuracy is a function not only of the eye-tracker itself but also of the state of the wearer. Future research could investigate the relative performance of this and other eye-trackers under a wider range of task conditions and participant samples.

Acknowledgments

This research is funded by grant 016.Vidi.178.047 ("How should automated vehicles communicate with other road users?"), which is provided by the Netherlands Organization for Scientific Research (NWO). We would like to thank Lokkeshver T. K. for his help during pilot studies conducted prior to the experiment.

Appendix 3.A. Results for Eye-Tracker Precision

The precision of the gaze for a participant in an experimental condition was calculated by computing the standard deviation of the angular distance θ over the time interval (10 seconds), and subsequently averaging the intervals of that participant and condition. Table 3.A1 shows the results of precision.

According to a two-way repeated-measures ANOVA, there was a significant effect of the Tobii model, $F(1,32) = 10.3$, $p = 0.003$, partial $\eta^2 = 0.24$, as well as of the level of dynamicity, $F(2,64) = 6.42$, $p = 0.003$, partial $\eta^2 = 0.17$, but there was no significant Tobii model × dynamicity interaction, $F(2,64) = 1.82$, $p = 0.171$, partial $\eta^2 = 0.05$.

Table 3.A1

Precision per experimental condition (in degrees). The mean, standard deviation (SD), and median across participants are reported.

Dynamicity condition	Eye-tracker	Mean	*SD*	Median	*n*
Seated, with chinrest	Tobii 2	1.22	0.77	1.00	36
Seated, without chinrest	Tobii 2	0.89	0.52	0.81	36
Walking	Tobii 2	1.29	1.01	0.92	36
Seated, with chinrest	Tobii 3	0.81	0.72	0.58	34
Seated, without chinrest	Tobii 3	0.77	0.82	0.45	33
Walking	Tobii 3	0.97	0.90	0.68	34

Appendix 3.B. Results for Eye-Tracker Bias

The bias of the gaze for a participant in an experimental condition was calculated by computing the mean gaze point and target point in pixels over the time interval (10 seconds), computing the angular distance of those means using Eq. 1, and subsequently averaging the intervals of that participant and condition.

Table 3.A2 shows the results of bias. According to a two-way repeated-measures ANOVA, there was a significant effect of the Tobii model, $F(1,32) = 32.0$, $p < 0.001$, partial $\eta^2 = 0.50$, a significant effect of dynamicity, $F(2, 64) = 4.60$, $p = 0.014$, partial $\eta^2 = 0.13$. Moreover, there was a significant Tobii model × dynamicity interaction, $F(2, 64) = 5.53$, $p = 0.006$, partial $\eta^2 = 0.15$.

Table 3.A2

Bias per experimental condition (in degrees). The mean, standard deviation (SD), and median across participants are reported.

Dynamicity condition	Eye-tracker	Mean	*SD*	Median	*n*
Seated, with chinrest	Tobii 2	2.22	1.10	1.97	36
Seated, without chinrest	Tobii 2	1.60	1.26	1.13	36
Walking	Tobii 2	2.86	2.01	2.52	36
Seated, with chinrest	Tobii 3	1.52	0.90	1.35	34
Seated, without chinrest	Tobii 3	1.30	0.71	1.10	33
Walking	Tobii 3	1.37	0.85	1.26	34

Supplementary Data

Data and scripts are available at:
https://doi.org/10.4121/442018c6-30eb-4439-a452-c0046726905c.

References

Aziz, S., & Komogortsev, O. (2022). An assessment of the eye tracking signal quality captured in the hololens 2. *Proceedings of the 2022 Symposium on Eye Tracking Research and Applications*, Seattle, WA, Article 5. https://doi.org/10.1145/3517031.3529626

Bahill, A. T., Adler, D., & Stark, L. (1975). Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology*, *14*, 468–469. https://iovs.arvojournals.org/arvo/content_public/journal/iovs/933061/468.pdf

Cercenelli, L., Tiberi, G., Bortolani, B., Giannaccare, G., Fresina, M., Campos, E., & Marcelli, E. (2019). Gaze Trajectory Index (GTI): A novel metric to quantify saccade trajectory deviation using eye tracking. *Computers in Biology and Medicine*, *107*, 86–96. https://doi.org/10.1016/j.compbiomed.2019.02.003

De Winter, J. C. F., Dodou, D., & Tabone, W. (2022). How do people distribute their attention while observing The Night Watch? *Perception, 51*, 763–788. https://doi.org/10.1177/03010066221122697

Dietrich, H., & Wuehr, M. (2019). Strategies for gaze stabilization critically depend on locomotor speed. *Neuroscience*, *408*, 418–429. https://doi.org/10.1016/j.neuroscience.2019.01.025

Ehinger, B. V., Groß, K., Ibs, I., & König, P. (2019). A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. *PeerJ*, *7*, Article e7086. https://doi.org/10.7717/peerj.7086

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*, 1920–1931. https://doi.org/10.1016/j.visres.2011.07.002

Franchak, J. M., McGee, B., & Blanch, G. (2021). Adapting the coordination of eyes and head to differences in task and environment during fully-mobile visual exploration. *PLOS ONE*, *16*, Article e0256463. https://doi.org/10.1371/journal.pone.0256463

Gibaldi, A., Vanegas, M., Bex, P. J., & Maiello, G. (2017). Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behavior Research Methods*, *49*, 923–946. https://doi.org/10.3758/s13428-016-0762-9

Holmqvist, K. (2017). *Common predictors of accuracy, precision and data loss in 12 eye-trackers.* ResearchGate. https://doi.org/10.13140/RG.2.2.16805.22246

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye-tracker hardware and its properties. In *Eye tracking: A comprehensive guide to methods and measures* (pp. 9–64). OUP Oxford.

Holmqvist, K., Örbom, S. L., Hooge, I. T. C., Niehorster, D. C., Alexander, R. G., Andersson, R., Benjamins, J. S., Blignaut, P., Brouwer, A.-M., Chuang, L. L., Dalrymple, K. A., Drieghe, D., Dunn, M. J., Ettinger, U., Fiedler, S., Foulsham, T., Van der Geest, J. N., Hansen, D. W., Hutton, S. B., ... Hessels, R. S. (2022). Eye tracking: Empirical foundations for a minimal reporting guideline. *Behavior Research Methods*, *55*, 364–416. https://doi.org/10.3758/s13428-021-01762-8

Hooge, I. T. C., Niehorster, D. C., Hessels, R. S., Benjamins, J. S., & Nyström, M. (2022). How robust are wearable eye trackers to slow and fast head and body movements? *Behavior Research Methods*, *55*, 4128–4142. https://doi.org/10.3758/s13428-022-02010-3

Hooge, I. T. C., Niehorster, D. C., Hessels, R. S., Cleveland, D., & Nyström, M. (2021). The pupil-size artefact (PSA) across time, viewing direction, and different

eye trackers. *Behavior Research Methods*, *53*, 1986–2006. https://doi.org/10.3758/s13428-020-01512-2

Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2012). Conveying clinical reasoning based on visual observation via eye-movement modelling examples. *Instructional Science*, *40*, 813–827. https://doi.org/10.1007/s11251-012-9218-5

Kredel, R., Vater, C., Klostermann, A., & Hossner, E.-J. (2017). Eye-tracking technology and the dynamics of natural gaze behavior in sports: A systematic review of 40 years of research. *Frontiers in Psychology*, *8*, Article 1845. https://doi.org/10.3389/fpsyg.2017.01845

Lamare, M. (1892). Des mouvements des yeux dans la lecture [Eye movements in reading]. *Bulletins et Mémoires de la Société Française d'Ophthalmologie*, *10*, 354–364. https://pure.mpg.de/rest/items/item_2352590/component/file_2628648/content

Lång, K., Zackrisson, S., Holmqvist, K., Nystrom, M., Andersson, I., Förnvik, D., Tingberg, A., & Timberg, P. (2011). Optimizing viewing procedures of breast tomosynthesis image volumes using eye tracking combined with a free response human observer study. *Proceedings of the Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment*, Orlando, FL, 15–25. https://doi.org/10.1117/12.878066

MacInnes, J. J., Iqbal, S., Pearson, J., & Johnson, E. N. (2018). *Wearable eye-tracking for research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices.* BioRxiv. https://doi.org/10.1101/299925

Mantiuk, R. (2017). Accuracy of high-end and self-build eye-tracking systems. In S. Kobayashi, A. Piegat, J. Pejaś, I. El Fray, & J. Kacprzyk (Eds.), *Hard and Soft Computing for Artificial Intelligence, Multimedia and Security. ACS 2016. Advances in Intelligent Systems and Computing* (pp. 216–227). Springer. https://doi.org/10.1007/978-3-319-48429-7_20

Meißner, M., Pfeiffer, J., Pfeiffer, T., & Oppewal, H. (2019). Combining virtual reality and mobile eye tracking to provide a naturalistic experimental environment for shopper research. *Journal of Business Research*, *100*, 445–458. https://doi.org/10.1016/j.jbusres.2017.09.028

Minakata, K., & Beier, S. (2021). The effect of font width on eye movements during reading. *Applied Ergonomics*, *97*, Article 103523. https://doi.org/10.1016/j.apergo.2021.103523

Moore, S. T., Hirasaki, E., Cohen, B., & Raphan, T. (1999). Effect of viewing distance on the generation of vertical eye movements during locomotion. *Experimental Brain Research*, *129*, 347–361. https://doi.org/10.1007/s002210050903

Morgante, J. D., Zolfaghari, R., & Johnson, S. P. (2012). A critical test of temporal and spatial accuracy of the Tobii T60XL eye tracker. *Infancy*, *17*, 9–32. https://doi.org/10.1111/j.1532-7078.2011.00089.x

Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2018). What to expect from your remote eye-tracker when participants are

unrestrained. *Behavior Research Methods*, *50*, 213–227. https://doi.org/10.3758/s13428-017-0863-0

Niehorster, D. C., Santini, T., Hessels, R. S., Hooge, I. T. C., Kasneci, E., & Nyström, M. (2020). The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods*, *52*, 1140–1160. https://doi.org/10.3758/s13428-019-01307-0

Nyström, M., Andersson, R., Holmqvist, K., & Van De Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, *45*, 272–288. https://doi.org/10.3758/s13428-012-0247-4

Onkhar, V., Bazilinskyy, P., Stapel, J. C. J., Dodou, D., Gavrila, D., & De Winter, J. C. F. (2021). Towards the detection of driver-pedestrian eye contact. *Pervasive and Mobile Computing, 76*, Article 101455. https://doi.org/10.1016/j.pmcj.2021.101455

Pastel, S., Chen, C.-H., Martin, L., Naujoks, M., Petri, K., & Witte, K. (2021). Comparison of gaze accuracy and precision in real-world and virtual reality. *Virtual Reality*, *25*, 175–189. https://doi.org/10.1007/s10055-020-00449-3

Płużyczka, M. (2018). The first hundred years: A history of eye tracking as a research method. *Applied Linguistics Papers*, 25, 101–116. https://doi.org/10.32612/uw.25449354.2018.4.pp.101-116

RateFast. (2015). RateFast Goniometer. https://blog.rate-fast.com/ratefast-goniometer

Rogers, S. L., Speelman, C. P., Guidetti, O., & Longmuir, M. (2018). Using dual eye tracking to uncover personal gaze patterns during social interaction. *Scientific Reports*, *8*, Article 4271. https://doi.org/10.1038/s41598-018-22726-7

Rosenberg, R., & Klein, C. (2015). The moving eye of the beholder: Eye tracking and the perception of paintings. In J. P. Huston, M. Nadal, F. Mora, L. F. Agnati, & C. J. Cela-Conde (Eds.), *Art, aesthetics and the brain* (pp. 79–108). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199670000.003.0005

Serchi, V., Peruzzi, A., Cereatti, A., & Della Croce, U. (2014). Tracking gaze while walking on a treadmill: Spatial accuracy and limits of use of a stationary remote eye-tracker. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, 3727–3730. https://doi.org/10.1109/EMBC.2014.6944433

Stahl, J. S. (1999). Amplitude of human head movements associated with horizontal saccades. *Experimental Brain Research*, *126*, 41–54. https://doi.org/10.1007/s002210050715

Stuart, S., Alcock, L., Godfrey, A., Lord, S., Rochester, L., & Galna, B. (2016). Accuracy and re-test reliability of mobile eye-tracking in Parkinson's disease and older adults. *Medical Engineering & Physics*, *38*, 308–315. https://doi.org/10.1016/j.medengphy.2015.12.001

't Hart, B. M., & Einhäuser, W. (2012). Mind the step: Complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Experimental Brain Research*, *223*, 233–249. https://doi.org/10.1007/s00221-012-3254-x

Thibeault, M., Jesteen, M., & Beitman, A. (2019). Improved accuracy test method for mobile eye tracking in usability scenarios. *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting, *63*, 2226–2230. https://doi.org/10.1177/1071181319631083

Tobii AB. (2017a). Tobii Pro Glasses 2. User manual. https://www.manualslib.com/download/1269253/Tobii-Pro-Glasses-2.html

Tobii AB. (2017b). Eye tracker data quality report: Accuracy, precision and detected gaze under optimal conditions—controlled environment, Tobii Pro Glasses 2.

Tobii AB. (2022a). Tobii Pro Glasses 3. User manual. https://go.tobii.com/tobii-pro-glasses-3-user-manual

Tobii AB. (2022b). Tobii Pro Glasses 3 data quality test report: Accuracy, precision, and data loss under controlled environment (Rev. 1).

Wang, B., & Grossman, T. (2020). BlyncSync: Enabling multimodal smartwatch gestures with synchronous touch and blink. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI. https://doi.org/10.1145/3313831.3376132

Werner, K., Raab, M., & Fischer, M. H. (2019). Moving arms: The effects of sensorimotor information on the problem-solving process. *Thinking & Reasoning*, *25*, 171–191. https://doi.org/10.1080/13546783.2018.1494630

Wyatt, H. J. (2010). The human pupil and the use of video-based eyetrackers. *Vision Research*, *50*, 1982–1988. https://doi.org/10.1016/j.visres.2010.07.008

Xia, Y., Liang, J., Li, Q., Xin, P., & Zhang, N. (2022). High-accuracy 3D gaze estimation with efficient recalibration for head-mounted gaze tracking systems. *Sensors*, *22*, Article 4357. https://doi.org/10.3390/s22124357

Yuen, H. K., Princen, J., Illingworth, J., & Kittler, J. (1990). Comparative study of Hough transform methods for circle finding. *Image and Vision Computing, 8*, 71–77. https://doi.org/10.1016/0262-8856(90)90059-e

Chapter 4

Towards the detection of driver-pedestrian eye contact

This chapter has been published as:
Onkhar, V., Bazilinskyy, P., Stapel, J. C. J., Dodou, D., Gavrila, D., & De Winter, J. C. F. (2021). Towards the detection of driver–pedestrian eye contact. *Pervasive and Mobile Computing*, *76*, Article 101455. https://doi.org/10.1016/j.pmcj.2021.101455

Abstract

Non-verbal communication, such as eye contact between drivers and pedestrians, has been regarded as one way to reduce accident risk. So far, studies have assumed rather than objectively measured the occurrence of eye contact. We address this research gap by developing an eye contact detection method and testing it in an indoor experiment with scripted driver-pedestrian interactions at a pedestrian crossing. Thirty participants acted as a pedestrian either standing on an imaginary curb or crossing an imaginary one-lane road in front of a stationary vehicle with an experimenter in the driver's seat. In half of the trials, pedestrians were instructed to make eye contact with the driver; in the other half, they were prohibited from doing so. Both parties' gaze was recorded using eye-trackers. An in-vehicle stereo camera recorded the car's point of view, a head-mounted camera recorded the pedestrian's point of view, and the location of the driver's and pedestrian's eyes was estimated using image recognition. We demonstrate that eye contact can be detected by measuring the angles between the vector joining the estimated location of the driver's and pedestrian's eyes, and the pedestrian's and driver's instantaneous gaze directions, respectively, and identifying whether these angles fall below a threshold of 4°. We achieved 100% correct classification of the trials involving eye contact and those without eye contact, based on measured eye contact duration. The proposed eye contact detection method may be useful for future research into eye contact.

4.1. Introduction

In 2018, there were over 300,000 pedestrian deaths worldwide (World Health Organization, 2018). Studies have shown that pedestrian fatalities are growing by the year, especially on urban roads (National Highway Traffic Safety Administration, 2018a), and that most pedestrian casualties occur during the act of street crossing (DaSilva et al., 2004). An area of study with relevance to pedestrian safety is how pedestrians interact with approaching vehicles. Next to formal traffic rules, non-verbal communication plays a role in the safe interaction between pedestrians and drivers (Färber, 2016; Habibovic et al., 2018).

4.1.1. The Effect of Eye Contact on Pedestrians

Through interviews and on-site observations (Lee et al., 2021; Sucha et al., 2017) and recordings of natural driving scenes (Rasouli et al., 2017; Schmidt & Färber, 2009; Sucha et al., 2017), it has been shown that a sizeable percentage of pedestrians use eye contact to negotiate right of way when crossing the road. Additionally, studies have investigated pedestrians' responses to automated vehicles without a driver making eye contact (typically using a Wizard of Oz approach; Habibovic et al., 2018; Malmsten Lundgren et al., 2017; Rothenbücher et al., 2015). In particular, Malmsten Lundgren et al. found that most pedestrians were willing to cross the road when there was eye contact with the driver, whereas only a few were willing when the driver of the automated vehicle was inattentive. There also exists a prevalent belief outside academia that eye contact is of significance to the safety of

pedestrians, as evidenced by notices, signs, and advice issued by traffic safety organizations (london.ca, 2016; National Highway Traffic Safety Administration, 2018b; Veiligverkeer, 2020). The view that eye contact is important has even led to the development of anthropomorphic external human-machine interfaces (eHMIs) for automated vehicles. Chang et al. (2017), for example, tested a novel eHMI with dynamic eyes on the car.

While many studies have shown that pedestrians seek eye contact with drivers (Lee et al., 2021; Schneemann & Gohl, 2016; Sucha et al., 2017), it has been suggested that eye contact is not essential and that pedestrians often cross in front of vehicles by solely relying on vehicle motion cues (Dey & Terken, 2017; Lee et al., 2021; Moore et al., 2019). In an online study by AlAdawy et al. (2019), participants looking at photos of a car with a driver inside at different distances and under different lighting conditions reported that, in many situations, they could not even see the driver, let alone make eye contact.

4.1.2. The Effect of Eye Contact on Drivers

Next to pedestrians' communication needs at crossings, studies have investigated the effect of pedestrians' communication attempts, including eye contact, on drivers. As early as 1974, Snyder et al. noted in a field experiment on hitchhiking that drivers yielded more often when staged hitchhikers sought eye contact with them. Katz et al. (1975) found that drivers slowed down and yielded more often to pedestrians when the pedestrians initiated crossing but were *not* looking in the driver's direction, compared to when they were. More recently, in a field study measuring car speed profiles as a function of eye contact, Ren et al. (2016) found that drivers braked earlier for staged pedestrians who attempted to make eye contact than for those who did not. That said, Schmidt and Färber (2009) found that participants looking at videos of traffic scenes from a driver's perspective were able to make accurate predictions of pedestrians' crossing intentions even when the pedestrians' heads were occluded, suggesting that eye contact is not essential in traffic.

4.1.3. Literature Gap

From the above, there appears to be a need for further research into the importance of eye contact in traffic. However, as of present, no general conclusions can be obtained due to the variety of measurement methods employed. Measurements such as head orientation, as reported by experimenters standing on the roadside or recorded via cameras inside or outside of the vehicle, can be used to infer eye contact seeking (Kotseruba et al., 2016; Rasouli et al., 2017; Roth et al., 2016; Schneemann & Gohl, 2016; Sucha et al., 2017). For example, based on an analysis of video clips from urban driving scenarios, Rasouli et al. (2017) suggested that pedestrians looking in the direction of approaching vehicles for longer than 1 second might also seek eye contact. However, head orientation alone does not determine where road users are looking.

Herein, we propose the use of eye-tracking to detect eye contact. Eye-tracking can establish where road users look without explicitly asking them and without relying on third-party observations. Some research into how drivers look at pedestrians already exists. For example, Walker (2005) and Walker and Brosnan (2007) reported that drivers gazed at cyclists' faces first and for longer than other body parts. Nathanael et al. (2019) used eye-tracking to analyze drivers' gaze and concluded that pedestrians' body movement/posture and eye gaze were sufficient to resolve crossing conflicts in the majority of interactions, without the need for eye contact or hand gestures. Diederichs et al. (2015) performed eye-tracking in a driving simulator study with simulated pedestrians and reported that drivers' pedal responses that indicated an intention to brake were accompanied by eye fixations on pedestrians 0.4–2.4 s earlier. Finally, Borowsky et al. (2012) conducted an eye-tracking experiment where participants viewed traffic videos from a driver's perspective and pressed a button when they perceived a hazard. The authors reported that, in general, drivers fixated more often on pedestrians on the road compared to those on the curb.

Research on eye movements in pedestrians exists as well. For example, De Winter et al. (2021) conducted an eye-tracking study of pedestrians during interactions with vehicles in a parking lot and found that pedestrians frequently sought eye contact with drivers. Dey et al. (2019) used eye-tracking of pedestrians at a curb and measured their willingness to cross in front of an oncoming vehicle using a handheld slider. They reported that despite the interior of approaching cars being dark and reflections on the windshield making it difficult to establish eye contact, pedestrians still sought information about the driver's intentions by looking at the windshield when the vehicle was nearby. A recent literature review on eye-tracking studies of pedestrians crossing the road noted that a limitation of the research so far is that the eye-tracking results were not combined with physical measurements such as the distance of the vehicle to the pedestrian (Lévêque et al., 2020). Image recognition combined with eye-tracking could prove a solution to this limitation.

Another important caveat in the above studies involving eye-tracking in traffic (e.g., De Winter et al., 2021; Dey et al., 2019; Nathanael et al., 2019) is that only one perspective (either the driver's or the pedestrian's) was measured, which provides an incomplete picture because eye contact is a mutual phenomenon (and see Roth et al., 2016 for a study on mutual situation awareness of driver and pedestrian; also Broz et al., 2012 and Rogers et al., 2018, for studies using dual eye-tracking in social interaction). In other words, gaze detection of solely one of the two parties is informative about the party's *seeking* of eye contact but cannot tell whether eye contact has been established. These problems in the operationalization of eye contact in the literature have also been reported by Jongerius et al. (2020) in their scoping review on eye contact in human-human interaction. This vacuum in the literature could be filled by techniques that detect driver-pedestrian eye contact.

4.1.4. Study Aims

In the current study, we developed a method that detects driver-pedestrian eye contact by means of two eye-trackers along with two cameras. Our work's novelties are the use of dual eye-tracking in the traffic context, which pinpoints where both the driver and the pedestrian are looking at any given time, and the use of image recognition on video recordings from cameras to estimate the locations of the driver and the pedestrian. We validated the method by means of an indoor experiment with scripted driver-pedestrian interactions at a pedestrian crossing.

Driver-pedestrian eye contact was operationalized as a situation when the driver and the pedestrian are looking at each other at the same time within predefined threshold angles. In the literature, there is an emphasis on the psychological (and hence, subjective) experience of eye contact (Heron, 1970; Jongerius et al., 2020). In this study, we are concerned with the objective detection of eye contact. An underlying assumption in our operationalization is that if two persons are looking at each other's faces, they are looking at each other's eyes in an attempt to make eye contact.

4.2. Methods

4.2.1. Participants

Thirty-one persons (23 males, 8 females) took part in the experiment as staged pedestrians. Participants were recruited via social media and personal contacts. Only people with normal visual acuity or corrected with contact lenses were eligible to participate. All participants provided written informed consent. The research was approved by the Human Research Ethics Committee of the Delft University of Technology (reference number 865). One male participant was excluded because of a failure of one eye-tracker, resulting in a final sample of 30, with a mean age (*SD*) of 24.8 (2.3) years and with ages ranging from 19 to 31 years.

4.2.2. Equipment

A head-mounted Tobii Pro Glasses 2 eye-tracker was used to track and record the pedestrian's (i.e., participant's) gaze direction at 50 Hz. The 'Gaze Spot Meter' setting of the Tobii was turned on, as a result of which the exposure of the camera images was automatically adjusted based on where the participant looked. A head-mounted camera built into the Tobii recorded the pedestrian's view as a video at 25 frames per second, a field of view of 90^{o}, and a resolution of 1920×1080. Pedestrian gaze calibration was achieved using a card with a printed bull's-eye, and parallax error was corrected automatically by the manufacturer's software.

A dashboard-mounted Smart Eye Pro dx eye-tracker installed in a Toyota Prius intelligent vehicle (i.e., with environment-sensing capabilities; Ferranti et al., 2019) was used to track and record the driver's (i.e., experimenter's) gaze direction at 60 Hz. The Smart Eye collected gaze data using a combination of the manufacturer's bundled software and custom C++ programs running inside the Robot Operating System (ROS) on an Ubuntu Linux computer onboard the Toyota Prius. Parallax

error was avoided as the Smart Eye was not head-mounted and worked with 3D space rather than 2D projections on an image plane.

Both eye-trackers work on the principle of pupil corneal reflection, i.e., by using the angle between the locations of the pupil and reflections of infrared light on a person's cornea to determine their gaze direction (Tobii AB, 2021). To this end, artificial infrared light sources are employed, and the corneal reflections are captured by infrared cameras. The pedestrian's and the driver's gazes were both taken as the average of the gaze directions from each of their two eyes. An iDS UI-3060CP-C-HQ Rev. 2 stereo camera installed in the car with integrated pedestrian detection recorded the pedestrian's locations using the single-shot detection (SSD; Braun et al., 2019) technique at 10 Hz.

The Smart Eye was temporally synchronized with the stereo camera using Network Time Protocol (NTP) clients on a local area network (LAN) with a synchronization buffer, which minimized the difference in capture time between the two sources. The Smart Eye was further spatially calibrated relative to 22 reference points inside the vehicle, the locations of which were known (i.e., the position and orientation of the vehicle's sensors, including that of the stereo camera) obtained by laser scanning the Toyota Prius (methods described by Domhof et al., 2019). Driver gaze was calibrated using a different set of 7 known points in the vehicle's interior. Flashes from a NexTorch flashlight (300 lumens) during the experiment were captured by both the Tobii camera and the stereo camera to enable retrospective synchronization of the data from the two eye-trackers.

4.2.3. Experimental Procedure

The experiment was conducted indoors in an open lab space, with the area cordoned off to prevent interference from passers-by. The location was well lit by a combination of natural light (indirect and diffused through windows) and fluorescent tubes on the ceiling. The Toyota Prius was parked away from sunlight to reduce windshield glare and angled in such a way as to minimize the chance of unwanted detection of onlookers and passers-by by the stereo camera.

Participants played the role of a pedestrian, and two experimenters conducted the study: the first posed as the driver of the stationary car, and the second instructed the participant about their task and controlled the torch for the synchronization of the two eye-trackers. Upon arrival, participants read and signed the consent form, which also contained a brief overview of the experiment, its objectives, and their role in it. They subsequently completed a questionnaire on their height, age, sex, nationality, etc. Next, participants were provided with an oral explanation of how to interact with the car (to supplement what they had read). They were asked to wear the Tobii, which was calibrated using a card with a printed bull's-eye. In the meantime, the driver calibrated his gaze with the Smart Eye using a checkerboard. Participants were instructed to imagine that they were a pedestrian on a curb who had the

intention of crossing a one-lane road at an uncontrolled crossing when an approaching car had just slowed down to a stop.

Each participant performed six trials, as summarized in Table 4.1. The trials involved either a standing pedestrian (on the right or left imaginary curb) or a crossing pedestrian, with instructions to make/not make eye contact with the driver. Three repetitions were conducted per trial. Thus, each participant performed 18 repetitions in total (6 trials x 3 repetitions). The six trials were performed in two blocks: one containing all four standing trials and the other containing the two crossing trials. The order in which the blocks were performed and the order of the trials within the blocks were randomized. The repetitions within each trial were conducted back-to-back.

Table 4.1
Trials in the experiment

Pedestrian action/position	**Eye contact**	**Abbreviation**
Left standing pedestrian	Yes	L-S-EC
Left standing pedestrian	No	L-S-NEC
Right standing pedestrian	Yes	R-S-EC
Right standing pedestrian	No	R-S-NEC
Crossing pedestrian	Yes	C-EC
Crossing pedestrian	No	C-NEC

Before each trial, participants were instructed by the second experimenter about what type of trial would be performed next (i.e., whether/where to stand or cross, and whether to make/not make contact with the driver). In the standing trials, the pedestrian stood on the imaginary curb at a longitudinal distance of 5 m from the front of the car and at a lateral distance of 0.5 m from the side of the car, either to its left or right (see Figure 4.1). In the crossing trials, the pedestrian always started on the right curb.

In half of the standing trials, the pedestrians were asked to turn their head to briefly make eye contact with the driver, whereas in the other half, the pedestrians were asked to turn their head to briefly look at the body of the car (at a location of their preference) but refrain from gazing at the driver. In the crossing trials, the pedestrians were instructed to either maintain eye contact with the driver for the whole time as they walked towards the opposite curb across the imaginary road and back (as shown in Figure 4.2a) or avoid it by fixating on the body of the car. One repetition in the standing trials involved head turning to look at the car/driver, followed by eye contact/no eye contact, and head-turning to look away from the car/driver. One repetition in the crossing trials involved head turning to look at the car/driver, followed by walking once in front of the vehicle towards the opposite curb and back while maintaining/avoiding eye contact, and head-turning to look away from the car/driver.

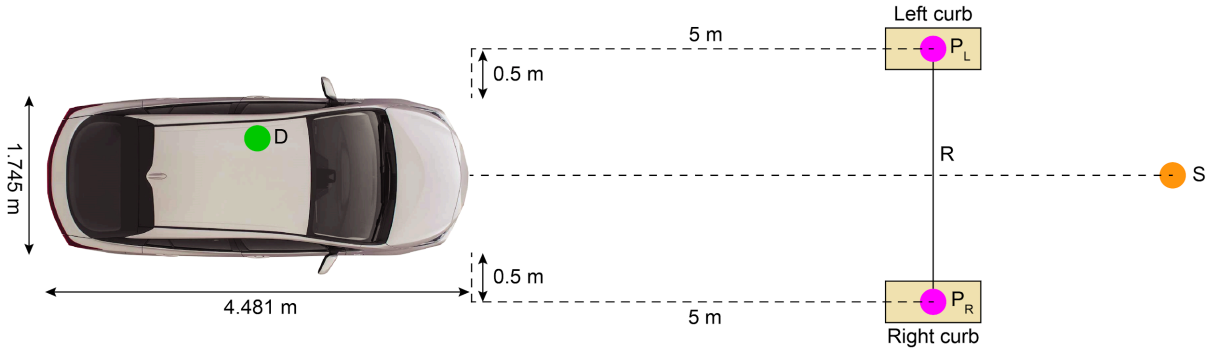

Figure 4.1. Layout of the experiment. The green circle marked 'D' denotes the location of the driver, the orange circle marked 'S' denotes the location of the 'synchronizer' (i.e., the experimenter with the torch), and the magenta circles marked 'P_L' and 'P_R' denote the locations of the pedestrian on the left and right curbs, respectively. In the experiment, the curbs were rectangles (shown here in light brown), and the road crossing (marked 'R') was a line, all three drawn in chalk on the floor.

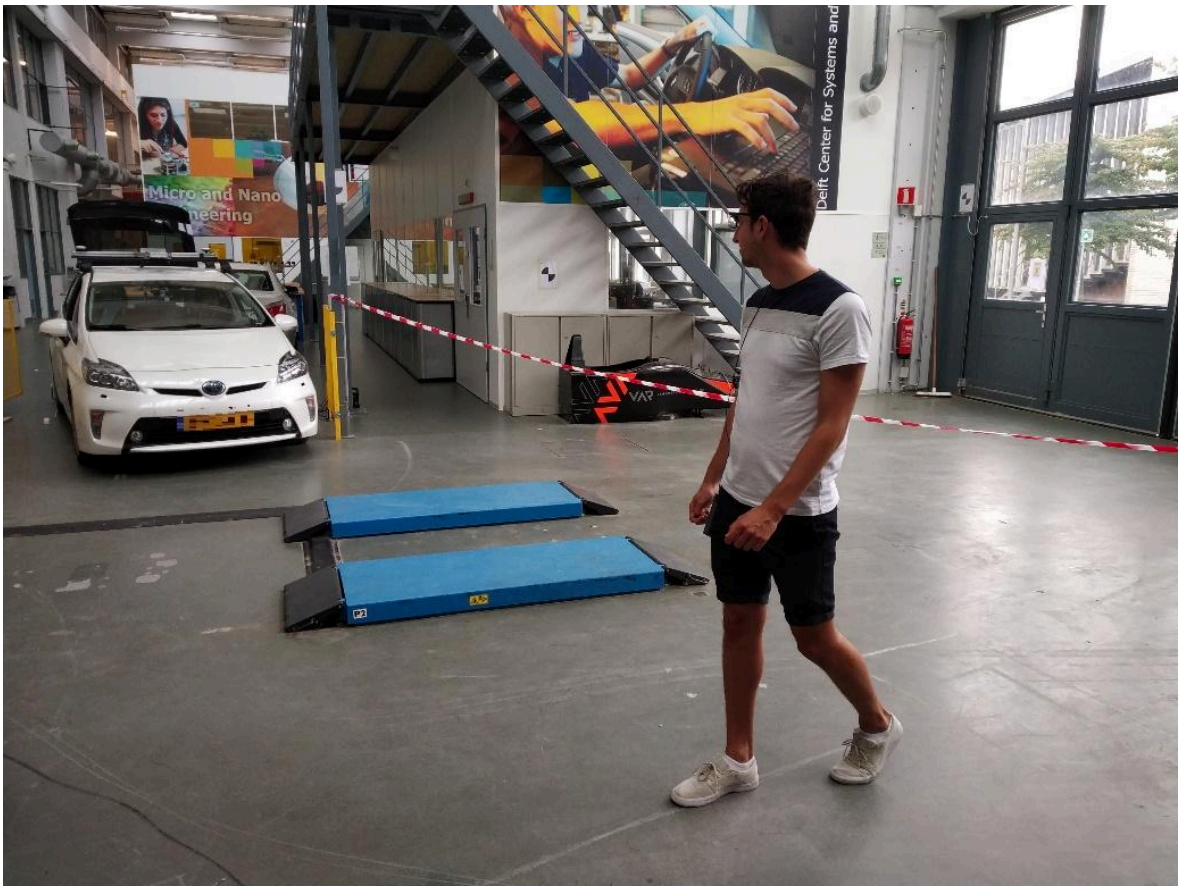

Figure 4.2a. A pedestrian crossing towards the right curb, with the driver and pedestrian looking at each other's eyes throughout the interaction. The participant provided permission for the publication of this image.

The driver briefly sought eye contact with the pedestrian in all standing trials, irrespective of whether the pedestrian was looking back at him. There was no predefined duration for the driver's eye contact seeking in the standing trials as this would be difficult to execute perfectly, so it was left to his discretion and what felt

natural. In the crossing trials, the driver followed the eyes of the walking pedestrian with his gaze for the entire duration (as seen in Figure 4.2b), irrespective of whether the pedestrian was looking back at him.

Figure 4.2b. The driver watching a pedestrian who is crossing towards the right curb, with the driver and pedestrian looking at each other's eyes throughout the interaction. Both persons provided permission for the publication of this image.

Before and after each trial, the participant was instructed to look at the torch held by the experimenter (standing in front of the car a few meters beyond the pedestrian) and wait for a flash, which served as an instantaneous marker for the start and end of that trial, respectively. The driver also knew to look at the torch at the beginning and the end of every trial.

All trials were recorded by both the Tobii and stereo cameras. After completing all the trials, the participants completed a questionnaire about their participation experience on 7-point Likert items (see Table 4.A1 in Appendix 4.A).

4.3. Analysis

4.3.1. Data Export

The gaze data collected by the Tobii, such as the pedestrian's gaze direction and gaze points in pixel coordinates corresponding to its camera video, were exported as Microsoft Excel files via the Tobii Pro Lab software. The data collected by the Smart

Eye and the stereo camera, such as the driver's gaze direction, position of his eyes, and the pedestrian's location, were saved in ROSBAG format and subsequently converted to comma-separated values (CSV) files via ROS.

4.3.2. Gaze Data Quality

First, the quality of the raw gaze data of the experimental trials was assessed using a gaze sample percentage, defined as the percentage of the trial duration for which the driver's or pedestrian's gaze direction could be measured. Any instances with missing data on either the driver's end or the pedestrian's end were filled using the previous non-missing entry.

4.3.3. Driver Eye Contact Seeking

To determine the vector connecting the driver's eyes to the pedestrian's eyes in 3D space (henceforth referred to as the 'ideal driver gaze'), the location of the driver's eyes (i.e., the midpoint between his right and left eye) was measured by the Smart Eye. The location of the pedestrian's eyes (i.e., the midpoint between the right and left eye) was estimated based on the pedestrian's location (*x, y*) obtained from the stereo camera and the pedestrian's eye height (*z*) calculated as 0.1 m below the pedestrian's self-reported height. Variation in the pedestrian's eye height due to their gait was assumed to be negligible. The stereo camera also detected the location of the experimenter holding the torch, which counted as false positives in the pedestrian detection. These readings were eliminated using a heuristically determined distance threshold of 7.3 m; that is, detections nearer than the threshold were the pedestrian and therefore retained, and those farther were the experimenter and therefore discarded.

A global coordinate system was defined, and the measurements from the Smart Eye (i.e., driver's instantaneous gaze direction and the position of his eyes) and the stereo camera (i.e., pedestrian's location) were converted to it from the devices' coordinate systems, as seen in Eq. 1. This equation describes the quaternion transformation of a coordinate system.

$$\left[0, x_{universal}, y_{universal}, z_{universal}\right] = \\ = [s,\ a,\ b,\ c] * \left(\left[0, x_{device}, y_{device}, z_{device}\right] - \left[0, k_x, k_y, k_z\right]\right)[s,\ a,\ b,\ c] \quad (1)$$

with * denoting the complex conjugate. To accomplish the coordinate transformation, translation of the measurements was performed using a translation matrix $\left[k_x, k_y, k_z\right]$, followed by rotation using a quaternion $[s, a, b, c]$, as seen above. Finally, all data were resampled to a common sampling rate of 100 Hz using linear interpolation.

The driver's eye contact seeking for each sampling instant was determined using the angle between the 'ideal driver gaze' vector and the driver's instantaneous gaze direction vector, as shown in Eq. 2. This equation uses Euclidean geometry to find

the angle between two vectors in 3D that originate from a common point. Driver eye contact seeking was operationalized as gaze angle error θ_{driver} < 4°. This threshold was based on a visual inspection of the distribution of θ_{driver} for all eye contact (EC) trials combined.

$$\theta_{driver} = \left(\frac{\|ideal\ driver\ gaze \times instantaneous\ gaze\ direction\|}{ideal\ driver\ gaze \cdot instantaneous\ gaze\ direction}\right) \quad (2)$$

4.3.4. Pedestrian Eye Contact Seeking

The location of the driver's eyes in the video recordings from the pedestrian's head-mounted camera was estimated on a frame-by-frame basis using computer vision. First, for increasing computational speed, each frame was resized from 1920×1080 to 192×108 pixels, and only the blue dimension from the RGB dimensions was retained. Next, the location of the Toyota Prius in each frame was estimated using a normalized two-dimensional cross-correlation technique (i.e., template matching; cf. Briechle & Hanebeck, 2001) with 82 reference images of the car (also consisting of only the blue dimension) with a resolution of 51×36 pixels. These reference images were cropped from Tobii recordings when the car was in view under different lighting conditions, angles, and perspectives. Pixel coordinates marking the driver's eyes in the reference images were manually coded. The location of the driver's eyes in each frame was estimated by finding the reference image that offered the maximum correlation (while also being above an empirically determined correlation threshold of 0.73), as shown in Figure 4.3.

The pedestrian's gaze point in pixel coordinates in each frame of a recording was available from the Tobii gaze data, and these values were divided by ten to suit the resized frames of 192×108 pixels. Pedestrian eye contact seeking was operationalized as $\theta_{pedestrian}$ < 4°. $\theta_{pedestrian}$ was approximated by converting the Pythagorean distance in pixels between the location of the driver's eyes and the pedestrian's gaze point into an angle (Eq. 3). Since the width of the resized frames is 192 pixels and the field of view of the Tobii camera is 90°, an 8.53-pixel distance corresponds to an angle of 4°.[2]

$$\theta_{pedestrian} = \frac{90}{192}\sqrt{\begin{matrix}(driv.head\ location_x - ped.gaze\ point_x)^2 + \\ +\left(driv.head\ location_y - ped.gaze\ point_y\right)^2\end{matrix}} \quad (3)$$

[2] The approximation of 4° corresponding to 8.53 pixels relies on assumptions related to the Tobii's camera perspective. We discovered later on that the horizontal field of view of the Tobii camera is about 82°, not 90° (90°, reported by the manufacturer, is likely the diagonal field of view). At the same time, Eq (3) assumes small angles, which may be untenable. Using manual measurements with the Tobii, we found that at low eye eccentricities of −15° to 15° (i.e., looking ahead), a step of 4° corresponds to 8 to 9 pixels and that at very high eye eccentricities of −35° or 35° (i.e., eyes turned strongly towards the left or right), a step of 4° corresponds to about 12.0 pixels. It may be assumed that participants were mostly looking ahead and that our adopted threshold of 8.53 pixels is indeed close to 4°.

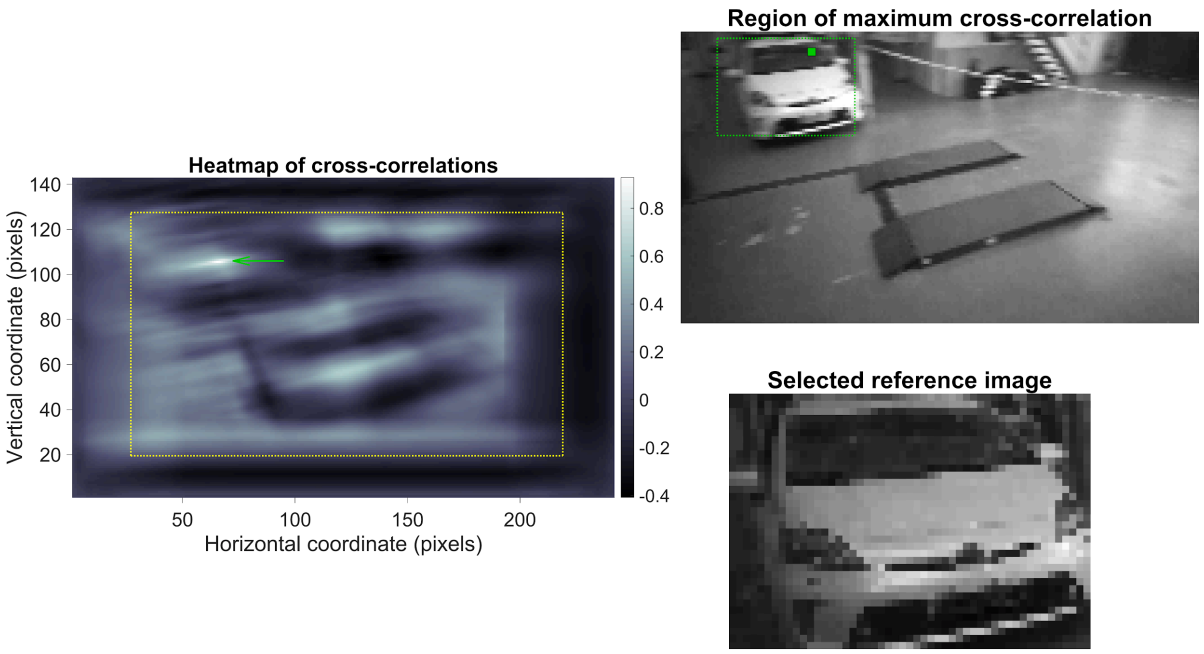

Figure 4.3. Heatmap of cross-correlations (left) between a resized Tobii video frame (top right) and the reference image having the highest value of maximum correlation with the frame among all reference images (bottom right). In the left figure, a green arrow marks the maximum correlation. A green box marks the region of maximum correlation. The driver's estimated position is marked with a green marker.

4.3.5. Driver-Pedestrian Eye Contact Establishment

To recap, driver eye contact seeking was operationalized as looking at the pedestrian with a gaze angle error smaller than 4°. Similarly, pedestrian eye contact seeking was operationalized as looking at the driver with a gaze angle error smaller than 4°. Eye contact was established when driver eye contact seeking and pedestrian eye contact seeking occurred concurrently. Finally, a classification between trials involving eye contact and those involving no eye contact was made by examining, per participant, whether the eye contact duration of the former type of trial exceeded the corresponding eye contact duration of the latter type of trial.

4.4. Results

4.4.1. Gaze Data Quality

Gaze data quality during the experiment was found to be high. The mean gaze sample percentage (SD) of the Smart Eye (i.e., driver) was 99.6% (0.7%) and 99.7% (0.6%) for the standing and crossing trials, respectively. The corresponding values of the Tobii (i.e., pedestrian) were 91.7% (5.3%) and 94.4% (4.4%) for the standing and crossing trials, respectively.

4.4.2. Driver Eye Contact Seeking

Figure 4.4 depicts the frequency distributions of driver gaze angle error (θ_{driver}) for all standing and crossing trials. A distinction is made between trials involving eye contact (EC), shown in blue, and trials involving no eye contact (NEC), shown as

black dotted lines. There is no marked difference in the curves in terms of the number of data points, which can be explained by the fact that the driver was not asked to adjust his gaze behavior based on the eye contact seeking behavior of the pedestrian. There was more eye contact in the crossing trials compared to the standing trials (see the ratio of the two peaks of the bimodal distribution), because the driver continuously tracked the pedestrian in the former case, whereas in the latter case, he only briefly sought eye contact with the pedestrian and looked away, and repeated this process thrice. From Figure 4.4, it can be seen that our threshold of 4° is a suitable choice, as it captures the large peak in the number of data points, which correspond to eye contact. More specifically, on combining the four distributions depicted in Figure 4.4, there were 2480 s of data for $0^o < \theta_{driver} \leq 4^o$ and only 60 s of data for $4^o < \theta_{driver} \leq 8^o$. When the driver was not looking at the pedestrian, he was either looking at the torch (at the start and end of every trial) or at the experimenter (after each repetition in the standing trials); this explains the second peak in the distributions, between 8° and 14°.

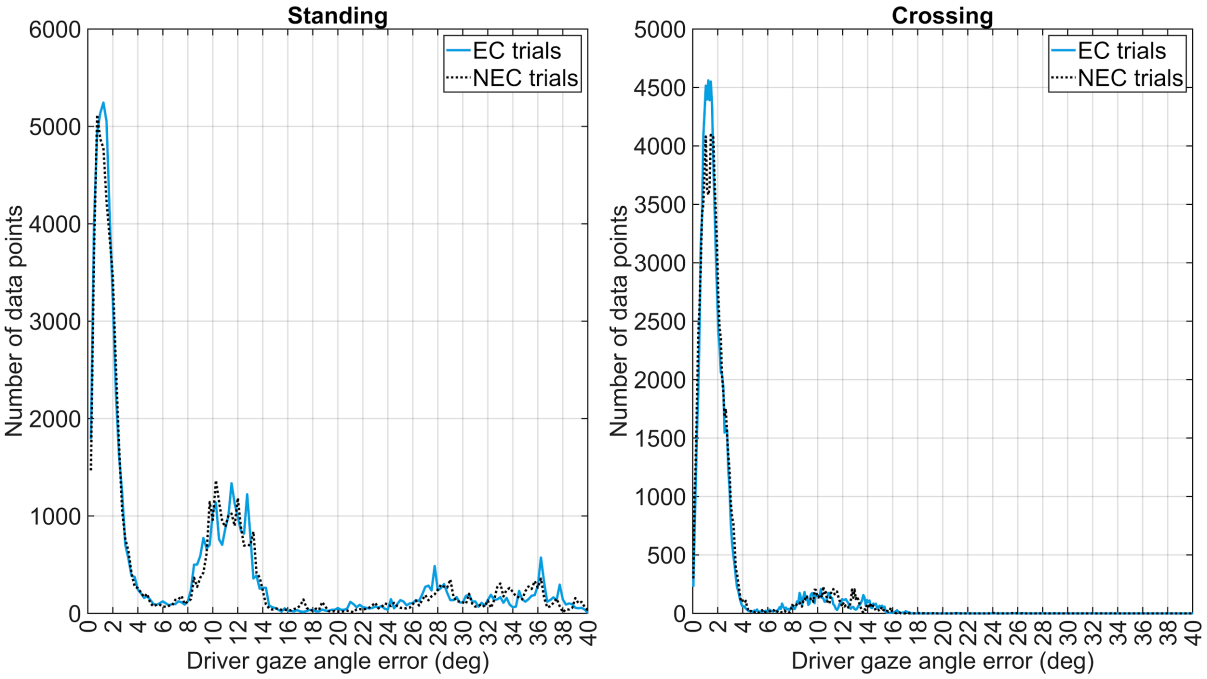

Figure 4.4. Distribution of the driver gaze angle error θ_{driver} (i.e., the angle between the instantaneous gaze direction of the driver and the 'ideal driver gaze') for eye contact (EC) and no eye contact (NEC) trials. The sampling rate was 100 Hz, and the resolution of the distribution was 0.25°. Left: Standing trials, Right: Crossing trials.

4.4.3. Pedestrian Eye Contact Seeking

Figure 4.5 shows similar information as Figure 4.4, but from the pedestrian's perspective. In contrast to the driver gaze angle error, the distributions of the pedestrian gaze angle error ($\theta_{pedestrian}$) are not bimodal but unimodal, which can be explained by individual differences, task instructions, and head and body rotation. That is, during the standing EC trials, participants were asked to seek eye contact

thrice and look away after each time, without instructions about where to look when looking away. Thus, due to varying levels of head and body rotation when looking away, the car (and therefore also the driver) was located at widely different parts of the visual field of the Tobii camera, or even completely outside of it. Accordingly, there are no distinct second peaks in the distribution of $\theta_{pedestrian}$, as was the case for driver gaze angle error shown in Figure 4.4. Figure 4.5 further shows a clear distinction in the pedestrian gaze angle error between the EC and NEC trials, with the mode of the distribution being about 1.5° (i.e., very close to the driver's eyes) for EC trials and about 12° (i.e., corresponding to some location on the car, as per the participants' task instructions) for NEC trials. The latter observation is corroborated by manual geometric calculations, which show that the pedestrian gaze angle error in an NEC trial when they are looking at the car's number plate (a plausible scenario) is 10–12°. This range of values is well above the 4° threshold, thereby reducing the likelihood of there being many false positives of eye contact in the NEC trials, assuming that the participants carried out the instructions faithfully. Our chosen $\theta_{pedestrian}$ threshold of 4° also represents a good trade-off for making a classification between EC and NEC trials, as gaze angle errors below this value capture a large portion of the samples in the EC trials (pedestrians were expected to exhibit small $\theta_{pedestrian}$ values for a portion of the EC trials) while capturing relatively few samples in the NEC trials (pedestrians were expected not to exhibit small $\theta_{pedestrian}$ values, since they were instructed not to make eye contact). Appendix 4.A provides further justification for the 4° threshold through a sensitivity analysis (see Figure 4.A1).

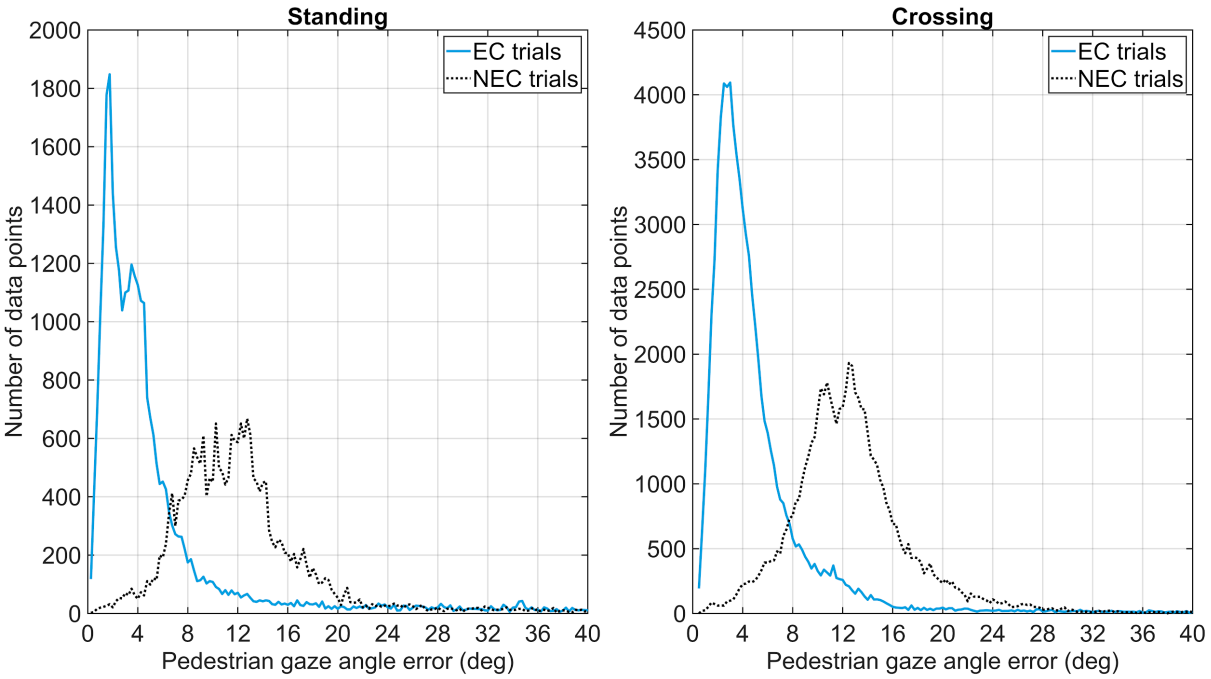

Figure 4.5. Distribution of the pedestrian gaze angle error $\theta_{pedestrian}$ (i.e., the angle between the gaze point of the pedestrian and the estimated position of the driver's eyes) for eye contact (EC) and no eye contact (NEC) trials. The sampling rate was 100 Hz, and the resolution of the distribution was 0.25°. Left: Standing trials, Right: Crossing trials.

4.4.4. Occurrence and Reconstruction of Eye Contact

Figure 4.6 shows four time-synchronized plots from a particular instant of a crossing trial with instructed eye contact (C-EC). In the bottom half, animations (comprising of a top view on the left and a side view on the right) of the trial are shown, created by plotting the position of the driver's eyes (green marker), his gaze direction (green dashed line), and the position of the pedestrian's eyes (magenta marker), along with an image of the Toyota Prius. Along the axes, *X*, *Y*, and *Z* represent the longitudinal, lateral, and vertical position in the world. The pedestrian's gaze direction is intentionally not plotted, as there was no sufficiently accurate way to translate it from a 2D pixel coordinate in the Tobii recordings to a 3D gaze direction for the animations. However, this omission does not have a bearing on detecting eye contact. The left top shows a Tobii camera screenshot with the pedestrian's view during a crossing trial, resized to a resolution of 192×108 pixels, and with the pedestrian's gaze point and the driver's estimated location overlaid as magenta and green markers, respectively. The top right part shows the driver's and pedestrian's gaze angle error plotted over time together with the 4° threshold of eye contact. Eye contact occurs when the values for the driver (green) and pedestrian (magenta) are both under the horizontal threshold line at the same time. It can be seen that for most of the trial, both the driver and the pedestrian sought eye contact.

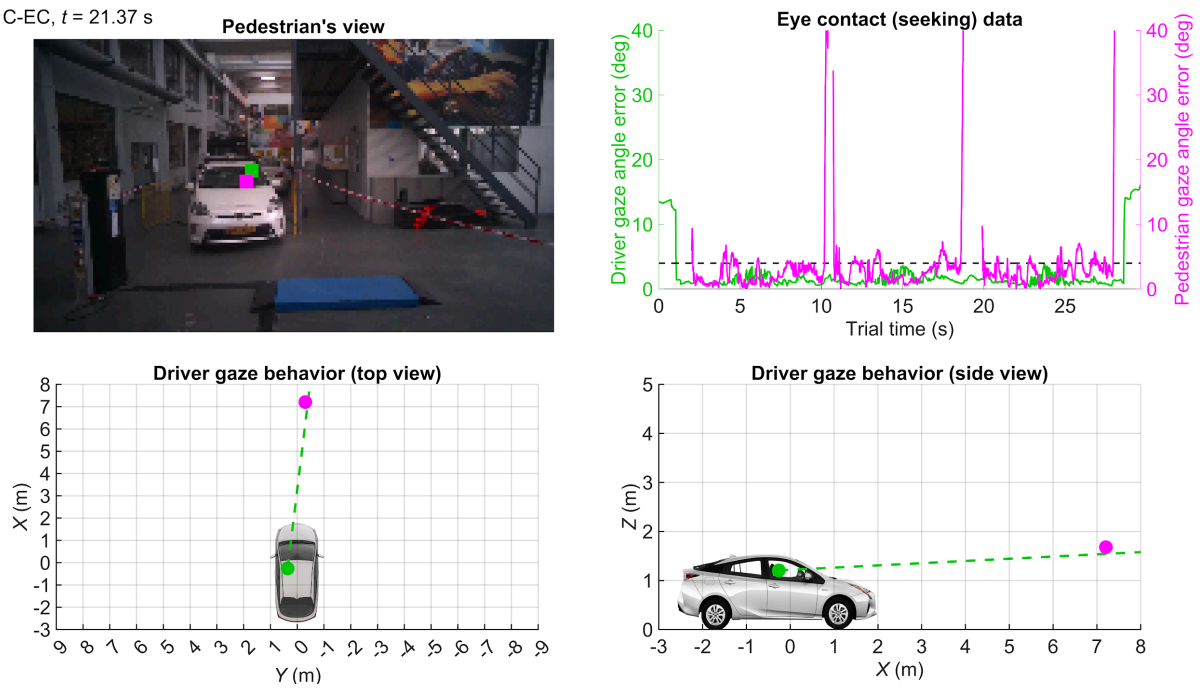

Figure 4.6. Eye contact overview of one trial (Crossing, Eye Contact; C-EC). Top left figure: Pedestrian's view. Top right figure: Driver and pedestrian gaze angle error as a function of time. Bottom left figure: Top view. Bottom right figure: Side view. Green marker: Estimated location of the driver's eyes. Magenta marker: Estimated location of the pedestrian's eyes (bottom figures) or pedestrian's gaze point (top left figure). Demo video available at: https://youtu.be/waDT32Tm-T4

The large gaze angle errors in the first and last two seconds of the trial are caused by the driver and pedestrian looking away from and towards the torch flash. The other sharp spikes at around 10 s and 19 s in the pedestrian's eye contact seeking graph are due to looking away from the driver upon completing a repetition (i.e., walking towards the opposite curb and back once). The gaps accompanying the spikes are due to the gaze angle error being undefined because the car (and therefore also the driver) being out of view in the Tobii camera.

4.4.5. Eye Contact Durations

Table 4.2 shows the means and standard deviations across the 30 participants for eye contact measures for the driver and the pedestrian. The driver sought eye contact about 55% of the time in the standing trials and in about 90% of the time in the crossing trials. Pedestrians sought eye contact for 20–25% of the duration of the standing trials with instructed eye contact and roughly 45% of the duration of the corresponding crossing trials. These values are consistent with the experimental protocol, where both parties sought intermittent eye contact with each other in the standing trials and visually tracked each other in the crossing trials. The pedestrian's percentages are lower than the driver's, as expected, because of the greater extent of head and body rotations by the pedestrian, thereby taking longer to (re-)establish eye contact.

Table 4.2

Means and standard deviations of eye contact measures for the standing trials and crossing trials. One trial consists of three repetitions.

	L-S-EC Mean	**L-S-NEC Mean**	**R-S-EC Mean**	**R-S-NEC Mean**	***L-S-EC SD***	***L-S-NEC SD***	***R-S-EC SD***	***R-S-NEC SD***
Trial duration (s)	12.35	11.97	12.38	11.81	3.25	2.90	2.72	2.71
Driver eye contact seeking (s)	6.82	6.59	6.80	6.47	2.18	1.92	1.63	1.79
Driver eye contact seeking (% of time)	55.0%	54.8%	55.1%	54.4%	5.6%	6.6%	5.9%	5.6%
Pedestrian eye contact seeking (s)	3.18	0.12	2.75	0.09	1.50	0.33	1.16	0.21
Pedestrian eye contact seeking (% of time)	26.0%	1.0%	22.6%	0.8%	11.1%	2.9%	9.2%	1.8%
Eye contact (s)	2.94	0.11	2.52	0.08	1.35	0.33	1.04	0.19
Eye contact (% of time)	24.3%	0.9%	20.8%	0.7%	10.7%	2.9%	8.7%	1.7%
	C-EC Mean	**C-NEC Mean**	**C-EC *SD***	**C-NEC *SD***				
Trial duration (s)	30.83	30.94	4.70	4.90				
Driver eye contact seeking (s)	27.94	28.05	4.11	4.41				
Driver eye contact seeking (% of time)	90.7%	90.7%	2.4%	3.1%				
Pedestrian eye contact seeking (s)	14.94	0.51	5.91	0.77				
Pedestrian eye contact seeking (% of time)	49.0%	1.7%	19.5%	2.6%				
Eye contact (s)	14.61	0.45	5.72	0.67				
Eye contact (% of time)	47.9%	1.5%	18.9%	2.2%				

Mean pedestrian eye contact seeking duration was 3.18 s in the L-S-EC trials, 2.75 s in the R-S-EC trials, and 14.94 s in the C-EC trials (see Table 4.2). Since each trial consisted of three repetitions, pedestrians sought eye contact for an average of 1.06 s, 0.92 s, and 4.98 s in a single repetition, respectively. Pedestrians crossed the road

twice in a single repetition, so the average pedestrian eye contact seeking duration in only one direction was half of 4.98 s, or 2.49 s.

The mean durations of eye contact were 2.94 s, 2.52 s, and 14.61 s for the L-S-EC, R-S-EC, and C-EC trials, respectively (see Table 4.2). This meant that in one repetition, eye contact lasted for 0.98 s, 0.84 s, and 4.87 s, in that order. Dividing the third value by two gives the average eye contact duration per repetition while crossing the road one-way, that is, 2.43 s.

4.4.6. Classification of Trials

Figure 4.7 illustrates the classification performance of trials with and without instructed eye contact based on eye contact duration. It can be seen that the type of trial is distinguished with 100% accuracy within-subject. In other words, all markers in Figure 4.7 lie below the diagonal line.

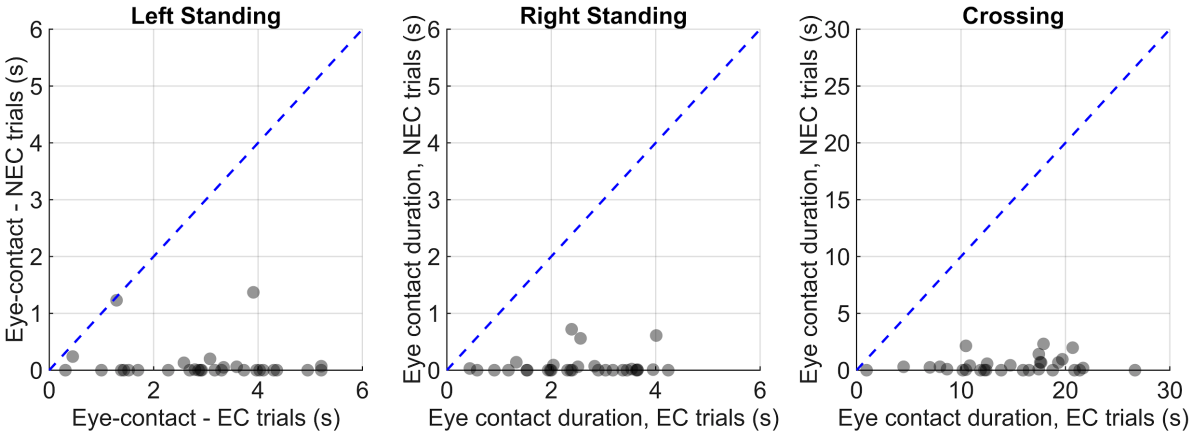

Figure 4.7. Comparison of pedestrian eye contact durations in the trials with and without instructed eye contact. Each marker represents one participant.

4.5. Discussion

4.5.1. Main Findings

This study aimed to develop an eye contact detection method to address the research gap in the objective measurement of eye contact in the traffic context. Our method's main innovation was the use of two eye-trackers to detect driver-pedestrian eye contact. The use of computer vision techniques to estimate the driver's and pedestrian's locations eliminates the need for manually coding the areas of interest. Compared to existing techniques such as self-reports and button press responses for recording eye contact, our dual eye-tracking method is accurate, since it does not rely on subjective perceptions of eye contact occurrence and is not influenced by reaction times. Accordingly, our setup may be useful for experimental research in staged scenarios and may form the first step towards real-time eye contact detection in crossing conflicts.

We provided a new operationalization for eye contact, namely that it occurred when the gaze angle errors of the driver and the pedestrian were both below 4°. The 4° threshold was determined heuristically based on the angle error distributions. However, the selected threshold also appears to have psychological significance: A study using animated faces by Gamer and Hecht (2007) showed that eye gaze around faces (at a distance of 5 m) was in the form of a cone of angular width up to approximately 8°, which translates to a gaze angle error of up to 4°.

Our method was validated using staged interactions with and without eye contact and yielded perfect within-subject classification. Furthermore, we generated animations of the driver-pedestrian interactions, demonstrating that traffic encounters could be reconstructed using only the information obtained from cameras and eye-trackers. Such a visualization could prove useful to enhance the situational awareness of occupants of the vehicle (see Chang et al., 2019 for a top-down display that enhances situational awareness in automated vehicles).

Participants were instructed to seek eye contact with the driver briefly in the standing trials (but were not told to look for a specific amount of time) and to walk in front of the car in the crossing trials (but were not told how fast to walk). It is worth noting that pedestrians have mostly been observed seeking eye contact when a vehicle is close to them and moving at low speeds or stopped (Dey et al., 2019), which resembles our experimental setting. However, strong conclusions about how long each party seeks eye contact or how long eye contact lasts in a real driver-pedestrian interaction cannot be made using our measurements. The observed eye contact durations (0.9 s and 2.4 s for a standing and crossing pedestrian, respectively) may be higher than what one might expect in real traffic. At pedestrian crossings in real traffic, road users look at various elements of the scene, including signs and road markings (Bichicchi et al., 2017), not just the other party's eyes. Furthermore, pedestrians may stop glancing at the car when it has become clear that the pedestrian can cross before the car (see Croft & Panchuk, 2018, for a similar phenomenon in pedestrian-pedestrian interaction).

As a corollary of our operationalization of eye contact (i.e., a logical AND of the driver's and pedestrian's gaze behavior), eye contact duration was always less than or equal to the lower of the two eye contact seeking durations. Of note, the mean eye contact durations for the trials were closer in magnitude to the pedestrian's mean durations of eye contact seeking than those of the driver's. This was probably due to the driver constantly tracking the pedestrian in the crossing trials, whereas the pedestrian had to turn around and look away.

4.5.2. Limitations

Our method has a few limitations. First, our operationalization of eye contact does not include the subjective awareness that eye contact is occurring. This could have been measured using a think-aloud method or with event recorders (as noted by

Jongerius et al., 2020) in the hands of both parties. It would be interesting to examine the association between objective and subjective driver-pedestrian eye contact, something that has become possible through our eye contact detection method. Pedestrians in the present study reported being highly involved in the task (see Table 4.A1 in Appendix 4.A), so close congruence between subjective and objective eye contact is expected. In more demanding scenarios, it may be the case that the driver and the pedestrian are objectively looking at each other but not subjectively aware that they are making eye contact, i.e., the 'looking but not seeing' phenomenon (White & Caird, 2010).

A second limitation is that aspects of synchronization, image recognition, and data processing were still performed manually. It is noted that our algorithm, though we ran it offline, processes data on a frame-by-frame basis (i.e., without forward-looking filters), and therefore can be made to run in real-time. The Smart Eye and the stereo camera already reported driver gaze and pedestrian location in real-time. If the Tobii and its camera could also be configured to do the same via the API provided by the manufacturer, real-time eye contact detection is a viable target. We are currently developing a real-time pedestrian feedback system based on this proposed setup, where auditory feedback is provided to the pedestrian depending on whether the pedestrian has or has not looked at the car.

A third limitation concerns the artificial setup of our experiment. Our study involved a staged indoor experiment with a stationary vehicle, not to mention that most pedestrians are not equipped with wearable eye-trackers (Tabone et al., 2021). These issues may be solved in the future with greater affordability of eye-trackers and the advent of smart glasses or eye-tracking contact lenses (Khaldi et al., 2020). Findings of a pilot study revealed that the Tobii performed poorly at tracking the user's gaze outdoors, presumably due to infrared radiation in sunlight (Tobii AB, 2020). The pilot test also found that the visibility of the driver was compromised because of windshield glare outside. Windshield glare appears to be a factor that prevents eye contact in traffic (AlAdawy et al., 2019), suggesting a need for synthetic eye contact detection, such as ours.

A final limitation is that our method used basic computer vision techniques for detecting the pedestrian and vehicle. Although road user detection worked reliably in our case, more sophisticated methods, which are now becoming available for less than $100 (e.g., Nvidia Jetson; Süzen et al., 2020), would be required to make our method work with a wider variety of road users. As a proof of concept, we applied recent object recognition software intended for real-time usage (YOLOv5, Cui, 2021) on one of our experimental videos captured with the Tobii camera. Results showed that the algorithm detected the target car, a car in the background, and persons outside the car, but not the experimenter in the car (see Figure 4.8). Although the detection was not as robust as our template matching approach (the target car was often labeled a truck, bus, or train), there clearly appears to be potential for real-time

usage in situations with multiple different vehicles. It is worth remembering that our image technique also does not detect the driver in the car, and hence the choice of image recognition algorithm does not largely affect the detection of eye contact.

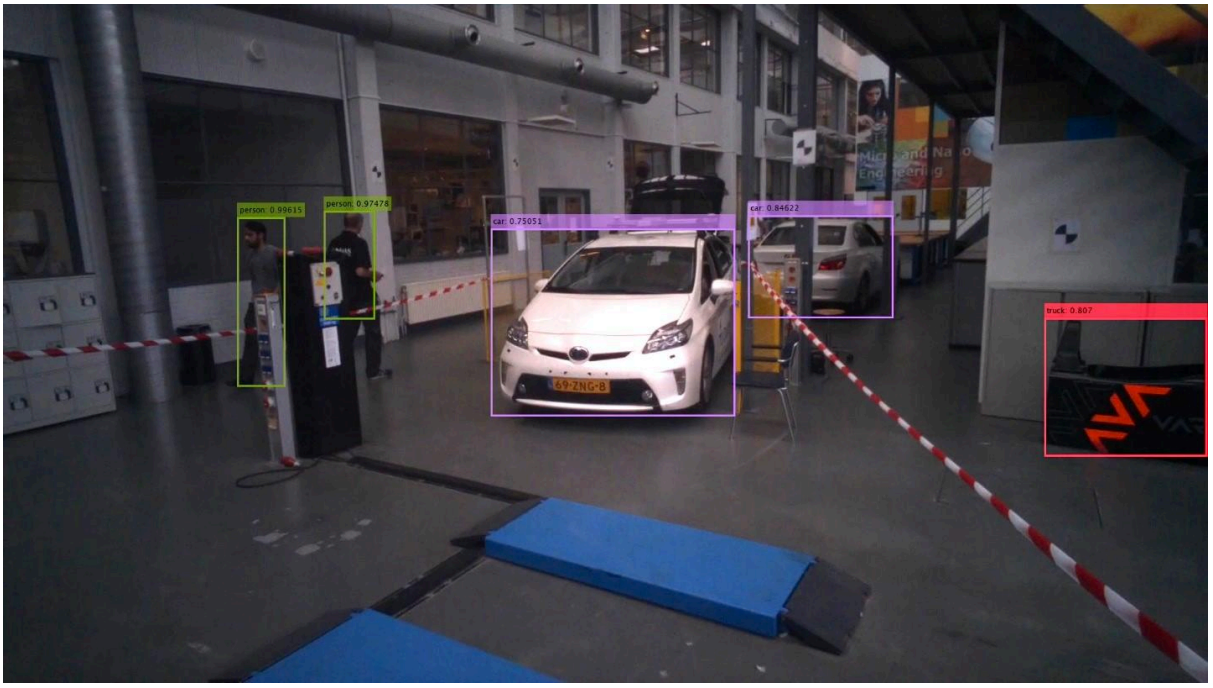

Figure 4.8. Screenshot of object recognition applied to one of our Tobii videos.

4.5.3. Outlook & Conclusion

There is ample scope for further research and applications. The topic of driver-pedestrian eye contact is not only of interest to manual and semi-automated driving (SAE Levels 0–2), it is also relevant to automated vehicles in which the driver is intermittently inattentive (SAE Levels 3 and 4). The vacuum created by missing eye contact in road interactions opens up possibilities to artificially substitute it. The anthropomorphic eye contact eHMI variants proposed by Ochiai and Toyoshima (2011), Chang et al. (2017), and Jaguar Land Rover (2018) are one way to achieve this. Eye contact could also be used as an objective input in vehicles or wearables for providing warnings (e.g., ‘mind the pedestrian’, ‘watch out, the driver is distracted’) or in automated vehicle control (e.g., braking earlier if there is no eye contact).

As pointed out above, an issue is that drivers and pedestrians are currently not equipped with eye-trackers. As an alternative to wearable eye-tracking, vehicle-based eye contact estimation may be possible through pedestrian head orientation estimation combined with contextual information (e.g., Quintero et al., 2014; Raza et al., 2018; Ridel et al., 2019; for a survey of methods, see Rudenko et al., 2020). However, for the time being, our method may be most useful for research purposes in staged scenarios. For example, our method could be applied to outdoor

experiments (in cloudy weather) to study eye contact in situations where right-of-way is not clear (e.g., Fu et al., 2019).

To conclude, the present study validated a novel eye contact detection method. Our method may stimulate further research that aims to obtain a deeper understanding of eye contact and its role in traffic.

Acknowledgement

This research is supported by grant 016.Vidi.178.047 (“How should automated vehicles communicate with other road users?”), which is financed by the Netherlands Organisation for Scientific Research (NWO). We want to express our gratitude to experimenters Lars Kooijman, Sparsh Sharma, Anand Sudha, and Arjun Anantharaman, who took turns to fulfill the role of using the torch in our study. An additional debt of thanks is owed to Lars Kooijman for his suggestions during the early stages of our work. We are also grateful to the participants for taking the time to help us conduct our research. We further appreciate the efforts of the many members (past and current) of the Department of Cognitive Robotics at the Delft University of Technology in building the intelligent vehicle we used in our experiment.

Appendix 4.A.

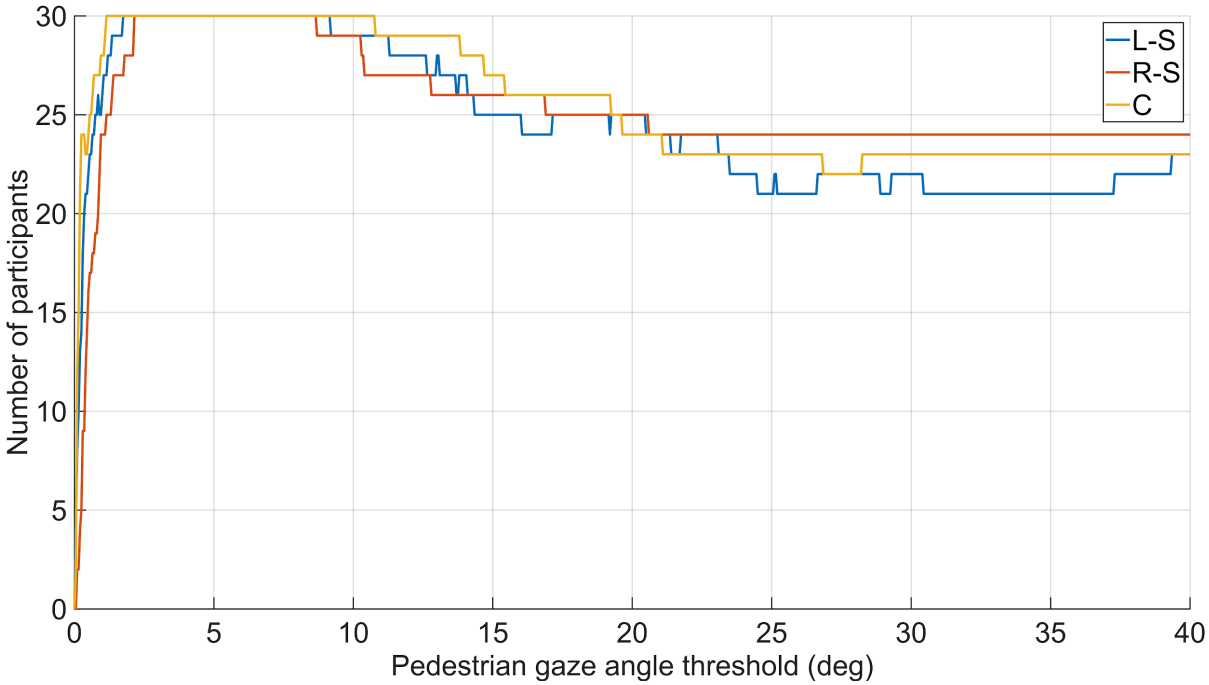

Figure 4.A1. Sensitivity analysis of the threshold for $\theta_{pedestrian}$. The *x*-axis shows the threshold angle (varied in 0.05° increments) and the *y*-axis shows the number of participants out of 30 for whom the number of data samples ($\theta_{pedestrian}$ < threshold) in the eye contact (EC) trial exceeded the number of data samples ($\theta_{pedestrian}$ < threshold) in the no eye contact (NEC) trial. Perfect classification is achieved for threshold angles between 1.15° and 10.75° for left standing (L-S) trials, between 2.15° and 8.65° for right standing (R-S) trials, and between 1.75° and 9.15° for crossing (C) trials.

Table 4.A1

Summary of responses to the post-experiment questionnaire. Responses were recorded on a 7-point Likert scale, with 1 denoting 'Not at all', and 7 denoting 'Completely'. Responses of 1–3 were considered negative, 4 neutral, and 5–7 positive.

Question	Positive response (% of participants)	Negative response (% of participants)	Neutral response (% of participants)	Mean response score (1 to 7)
How well were you able to imagine a real traffic scenario at a real pedestrian crossing?	64.5	19.3	16.2	4.74
How well could you concentrate on the task during the experiment?	100	0	0	6.42
How well could you maintain eye contact during the experiment without looking elsewhere?	90.3	0	9.7	5.87
How well could you avoid eye contact during the experiment and look elsewhere?	100	0	0	6.29
How conscious were you of wearing eye-tracking glasses while performing the experiment?	48.4	51.6	0	3.97
How much do you think the eye-tracking glasses affected your natural road crossing behavior?	22.6	64.5	12.9	2.87
How realistic did the experiment feel compared to crossing a real road?	45.2	22.5	32.3	4.32
How involved were you during the experiment?	96.8	0	3.2	6.32
How clear were the instructions for the experiment?	96.8	0	3.2	6.68
How well do you think you followed the instructions for the experiment?	96.8	0	3.2	6.48

Figure 4.A2. Screenshot from the stereo camera recording of a standing trial. The 'curbs' and the 'road' marked in chalk are visible here. Both persons provided permission for the publication of this image.

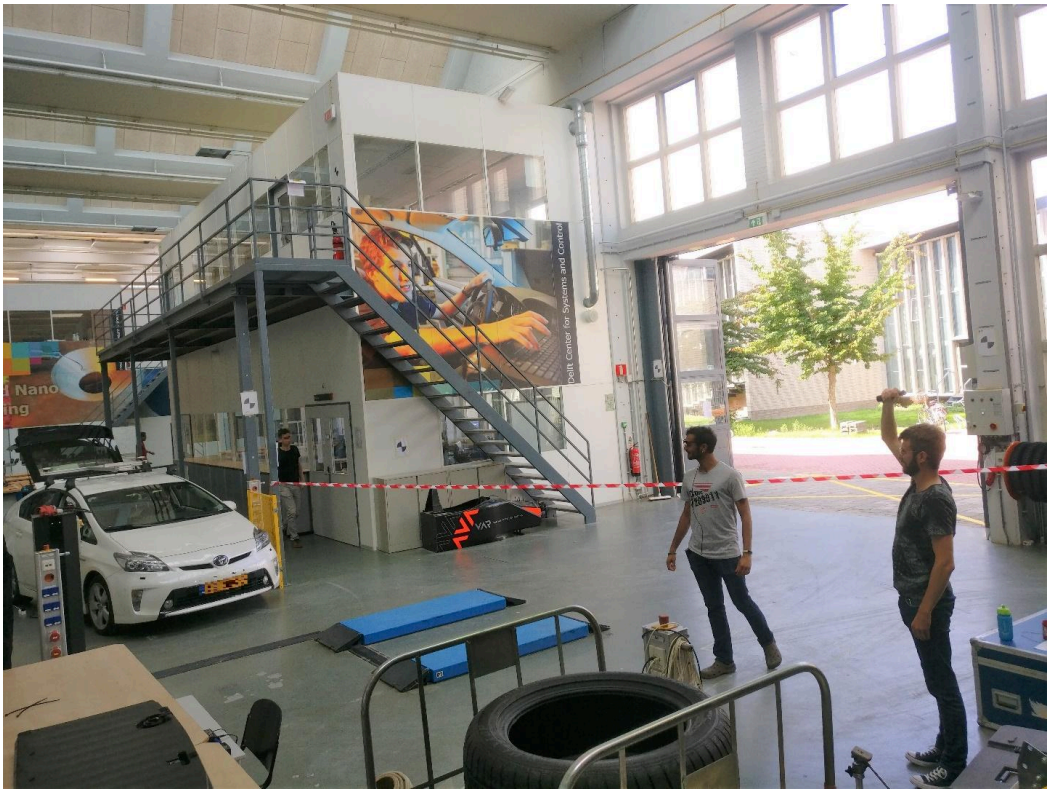

Figure 4.A3. A trial in progress, moments after eye-tracker synchronization, performed by the 'synchronizer' (far right) for the standing pedestrian on the left curb (second from right) and the car driver (obscured by windshield reflections in the photographer's point of view). No eye contact is occurring at this instant, with the driver looking at the pedestrian's eyes but the pedestrian looking at the right side-view mirror of the car. Both persons provided permission for the publication of this image.

Supplementary Data

The two questionnaires, a demo video corresponding to Figure 4.6, the data, and MATLAB codes used for the analyses, are available at: https://doi.org/10.4121/15134037.

References

AlAdawy, D., Glazer, M., Terwilliger, J., Schmidt, H., Domeyer, J., Mehler, B., Reimer, B., & Fridman, L. (2019). Eye contact between pedestrians and drivers. *Proceedings of the Tenth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Santa Fe, New Mexico, 301–307. https://pubs.lib.uiowa.edu/driving/article/28343/galley/136635/view

Bichicchi, A., Mazzotta, F., Lantieri, C., Vignali, V., Simone, A., Dondi, G., & Costa, M. (2017). The influence of pedestrian crossings features on driving behavior and road safety. In A. Bichicchi, F. Mazzotta, C. Lantieri, V. Vignali, A. Simone, G. Dondi, & M. Costa, *Transport Infrastructure and Systems* (pp. 741–746). CRC Press. https://doi.org/10.1201/9781315281896-95

Borowsky, A., Oron-Gilad, T., Meir, A., & Parmet, Y. (2012). Drivers' perception of vulnerable road users: A hazard perception approach. *Accident Analysis & Prevention*, 44, 160–166. https://doi.org/10.1016/j.aap.2010.11.029

Braun, M., Krebs, S., Flohr, F., & Gavrila, D. M. (2019). EuroCity Persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 1844–1861. https://doi.org/10.1109/TPAMI.2019.2897684

Briechle, K., & Hanebeck, U. D. (2001). Template matching using fast normalized cross correlation. *Optical Pattern Recognition XII*, *4387*, 95–102. https://doi.org/10.1117/12.421129

Broz, F., Lehmann, H., Nehaniv, C. L., & Dautenhahn, K. (2012). Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation. *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*, Paris, France, 858–864. https://doi.org/10.1109/ROMAN.2012.6343859

Chang, C. C., Grier, R. A., Maynard, J., Shutko, J., Blommer, M., Swaminathan, R., & Curry, R. (2019). Using a situational awareness display to improve rider trust and comfort with an AV taxi. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*, 2083–2087. https://doi.org/10.1177/1071181319631428

Chang, C.-M., Toda, K., Sakamoto, D., & Igarqashi, T. (2017). Eyes on a car: An interface design for communication between an autonomous car and a pedestrian. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Oldenburg, Germany, 65–73. https://doi.org/10.1145/3122986.3122989

Croft, J. L., & Panchuk, D. (2018). Watch where you're going? Interferer velocity and visual behavior predicts avoidance strategy during pedestrian encounters. *Journal of Motor Behavior*, 50, 353–363. https://doi.org/10.1080/00222895.2017.1363695

Cui. (2021). *yoloV5-matlab* [Computer software]. https://www.mathworks.com/matlabcentral/fileexchange/89012-yolov5-matlab

DaSilva, M. P., Smith, J. D., & Najm, W. G. (2004). *Analysis of pedestrian crashes* (Report No. DOT-VNTSC-NHTSA-02-02). Cambridge, MA: John A. Volpe National Transportation Systems Center. https://web.archive.org/web/20221005224445/http://www.nhtsa.gov/DOT/NHTSA/NRD/Multimedia/PDFs/Crash%20Avoidance/2003/DOTHS809585.pdf

De Winter, J. C. F., Bazilinskyy, P., Wesdorp, D., De Vlam, V., Hopmans, B., Visscher, J., & Dodou, D. (2021). How do pedestrians distribute their visual attention when walking through a parking garage? An eye-tracking study. *Ergonomics, 64*, 793–805. https://doi.org/10.1080/00140139.2020.1862310

Dey, D., & Terken, J. (2017). Pedestrian interaction with vehicles: Roles of explicit and implicit communication. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Oldenburg, Germany, 109–113. https://doi.org/10.1145/3122986.3123009

Dey, D., Walker, F., Martens, M., & Terken, J. (2019). Gaze patterns in pedestrian interaction with vehicles: Towards effective design of external human-machine interfaces for automated vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Utrecht, the Netherlands, 369–378. https://doi.org/10.1145/3342197.3344523

Diederichs, F., Schüttke, T., & Spath, D. (2015). Driver Intention Algorithm for Pedestrian Protection and Automated Emergency Braking Systems. *Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems*, Gran Canaria, Spain, 1049–1054. https://doi.org/10.1109/ITSC.2015.174

Domhof, J., Kooij, J. F. P., & Gavrila, D. M. (2019). An extrinsic calibration tool for radar, camera and lidar. *Proceedings of the International Conference on Robotics and Automation*, Montreal, Canada, 8107–8113. https://doi.org/10.1109/ICRA.2019.8794186

Färber, B. (2016). Communication and communication problems between autonomous vehicles and human drivers. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous driving* (pp. 125–144). Springer. https://doi.org/10.1007/978-3-662-48847-8_7

Ferranti, L., Brito, B., Pool, E., Zheng, Y., Ensing, R. M., Happee, R., Shyrokau, B., Kooij, J. F. P., Alonso-Mora, J., & Gavrila, D. M. (2019). SafeVRU: A research platform for the interaction of self-driving vehicles with vulnerable road users. *Proceedings of the 2019 IEEE Intelligent Vehicles Symposium*, Paris, France, 1660–1666. https://doi.org/10.1109/IVS.2019.8813899

Fu, T., Hu, W., Miranda-Moreno, L., & Saunier, N. (2019). Investigating secondary pedestrian-vehicle interactions at non-signalized intersections using vision-based trajectory data. *Transportation Research Part C: Emerging Technologies*, *105*, 222–240. https://doi.org/10.1016/j.trc.2019.06.001

Gamer, M., & Hecht, H. (2007). Are you looking at me? Measuring the cone of gaze. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 705–715. https://doi.org/10.1037/0096-1523.33.3.705

Habibovic, A., Malmsten Lundgren, V., Andersson, J., Klingegård, M., Lagström, T., Sirkka, A., Fagerlönn, J., Edgren, C., Fredriksson, R., Krupenia, S., Saluäär, D., & Larsson, P. (2018). Communicating intent of automated vehicles to pedestrians. *Frontiers in Psychology*, *9*, Article 1336. https://doi.org/10.3389/fpsyg.2018.01336

Heron, J. (1970). The phenomenology of social encounter: The gaze. *Philosophy and Phenomenological Research*, 31, 243–264. https://doi.org/10.2307/2105742

Jaguar Land Rover. (2018). The virtual eyes have it. https://web.archive.org/web/20181203150349/https://www.jaguarlandrover.com/2018/virtual-eyes-have-it

Jongerius, C., Hessels, R. S., Romijn, J. A., Smets, E. M. A., & Hillen, M. A. (2020). The measurement of eye contact in human interactions: A scoping review. *Journal of Nonverbal Behavior, 44*, 363–389. https://doi.org/10.1007/s10919-020-00333-3

Katz, A., Zaidel, D., & Elgrishi, A. (1975). An experimental study of driver and pedestrian interaction during the crossing conflict. *Human Factors*, *17*, 514–527. https://doi.org/10.1177/001872087501700510

Khaldi, A., Daniel, E., Massin, L., Kärnfelt, C., Ferranti, F., Lahuec, C., Seguin, F., Nourrit, V., & De Bougrenet de la Tocnaye, J. L. (2020). A laser emitting contact lens for eye tracking. *Scientific Reports*, *10*, Article 14804. https://doi.org/10.1038/s41598-020-71233-1

Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2016). *Joint attention in autonomous driving (JAAD).* arXiv. https://doi.org/10.48550/arXiv.1609.04741

Lee, Y. M., Madigan, R., Giles, O., Garach-Morcillo, L., Markkula, G., Fox, C., Camara, F., Rothmueller, M., Vendelbo-Larsen, S. A., Holm Rasmussen, P., Dietrich, A., Nathanael, D., Portouli, V., Schieben, A., & Merat, N. (2021). Road users rarely use explicit communication when interacting in today's traffic: Implications for automated vehicles. *Cognition, Technology & Work, 23*, 367–380. https://doi.org/10.1007/s10111-020-00635-y

Lévêque, L., Ranchet, M., Deniel, J., Bornard, J.-C., & Bellet, T. (2020). Where do pedestrians look when crossing? A state of the art of the eye-tracking studies. *IEEE Access, 8*, 164833–163843. https://doi.org/10.1109/ACCESS.2020.3021208

london.ca. (2016). Road danger reduction work programme. https://democracy.cityoflondon.gov.uk/documents/s78690/RoadDangerreduction.pdf

Malmsten Lundgren, V., Habibovic, A., Andersson, J., Lagström, T., Nilsson, M., Sirkka, A., Fagerlönn, J., Fredriksson, R., Edgren, C., Krupenia, S., & Saluäär, D. (2017). Will there be new communication needs when introducing automated vehicles to the urban context? In N. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation* (pp. 485–497). Springer. https://doi.org/10.1007/978-3-319-41682-3_41

Moore, D., Currano, R., Strack, G. E., & Sirkin, D. (2019). The case for implicit external human-machine interfaces for autonomous vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Utrecht, the Netherlands, 295–307. https://doi.org/10.1145/3342197.3345320

Nathanael, D., Portouli, E., Papakostopoulos, V., Gkikas, K., & Amditis, A. (2019). Naturalistic observation of interactions between car drivers and pedestrians in high density urban settings. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association* (pp. 389–397). Springer. https://doi.org/10.1007/978-3-319-96074-6_42

National Highway Traffic Safety Administration (NHTSA). (2018a). 2017 fatal motor vehicle crashes: Overview. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812603

National Highway Traffic Safety Administration (NHTSA). (2018b). Pedestrian safety. https://www.nhtsa.gov/road-safety/pedestrian-safety

Ochiai, Y., & Toyoshima, K. (2011). Homunculus: The vehicle as augmented clothes. *Proceedings of the 2nd Augmented Human International Conference*, Tokyo, Japan, Article 3. https://doi.org/10.1145/1959826.1959829

Quintero, R., Almeida, J., Llorca, D. F., & Sotelo, M. A. (2014). Pedestrian path prediction using body language traits. *The 2014 IEEE Intelligent Vehicles Symposium Proceedings*, Dearborn, MI, 317–323. https://doi.org/10.1109/IVS.2014.6856498

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Agreeing to cross: How drivers and pedestrians communicate. *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium*, Los Angeles, CA, 264–269. https://doi.org/10.1109/IVS.2017.7995730

Raza, M., Chen, Z., Rehman, S. U., Wang, P., & Bao, P. (2018). Appearance based pedestrians' head pose and body orientation estimation using deep learning. *Neurocomputing, 272*, 647–659. https://doi.org/10.1016/j.neucom.2017.07.029

Ren, Z., Jiang, X., & Wang, W. (2016). Analysis of the influence of pedestrians' eye contact on drivers' comfort boundary during the crossing conflict. *Procedia Engineering*, *137*, 399–406. https://doi.org/10.1016/j.proeng.2016.01.274

Ridel, D. A., Deo, N., Wolf, D., & Trivedi, M. (2019). Understanding pedestrian-vehicle interactions with vehicle mounted vision: An LSTM model and empirical analysis. *Proceedings of the 2019 IEEE Intelligent Vehicles Symposium*, Paris, France, 913–918. https://doi.org/10.1109/IVS.2019.8813798

Rogers, S. L., Speelman, C. P., Guidetti, O., & Longmuir, M. (2018). Using dual eye tracking to uncover personal gaze patterns during social interaction. *Scientific Reports, 8*, Article 4271. https://doi.org/10.1038/s41598-018-22726-7

Roth, M., Flohr, F., & Gavrila, D. M. (2016). Driver and pedestrian awareness-based collision risk analysis. *Proceedings of the 2016 IEEE Intelligent Vehicles Symposium*, Gothenburg, Sweden, 454–459. https://doi.org/10.1109/IVS.2016.7535425

Rothenbücher, D., Li, J., Sirkin, D., Mok, B., & Ju, W. (2015). Ghost driver: A platform for investigating interactions between pedestrians and driverless vehicles. *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Nottingham, UK, 44–49. https://doi.org/10.1145/2809730.2809755

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., & Arras, K. O. (2020). Human motion trajectory prediction: A survey. *The International Journal of Robotics Research, 39*, 895–935. https://doi.org/10.1177/0278364920917446

Schmidt, S., & Färber, B. (2009). Pedestrians at the kerb – Recognising the action intentions of humans. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*, 300–310. https://doi.org/10.1016/j.trf.2009.02.003

Schneemann, F., & Gohl, I. (2016). Analyzing driver-pedestrian interaction at crosswalks: A contribution to autonomous driving in urban environments. *Proceedings of the 2016 IEEE Intelligent Vehicles Symposium,* Gothenburg, Sweden, 38–43. https://doi.org/10.1109/IVS.2016.7535361

Snyder, M., Grather, J., & Keller, K. (1974). Staring and compliance: A field experiment on hitchhiking. *Journal of Applied Social Psychology*, *4*, 165–170. https://doi.org/10.1111/j.1559-1816.1974.tb00666.x

Sucha, M., Dostal, D., & Risser, R. (2017). Pedestrian-driver communication and decision strategies at marked crossings. *Accident Analysis & Prevention*, *102*, 41–50. https://doi.org/10.1016/j.aap.2017.02.018

Süzen, A. A., Duman, B., & Şen, B. (2020). Benchmark analysis of Jetson TX2, Jetson Nano and Raspberry PI using Deep-CNN. *Proceedings of the 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications*, Ankara, Turkey. https://doi.org/10.1109/HORA49412.2020.9152915

Tabone, W., De Winter, J. C. F., Ackermann, C., Bärgman, J., Baumann, M., Deb, S., Emmenegger, C., Habibovic, A., Hagenzieker, M., Hancock, P. A., Happee, R., Krems, J., Lee, J. D., Martens, M., Merat, N., Norman, D. A., Sheridan, T. B., & Stanton, N. A. (2021). Vulnerable road users and the coming wave of automated vehicles: Expert perspectives. *Transportation Research Interdisciplinary Perspectives*, *9*, Article 100293. https://doi.org/10.1016/j.trip.2020.100293

Tobii AB. (2020). Tobii Pro Glasses 2 User's manual, Version 1.1.3. https://web.archive.org/web/20200828092602/https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-glasses-2-user-manual.pdf

Tobii AB. (2021). How do Tobii eye trackers work? https://connect.tobii.com/s/article/How-do-Tobii-eye-trackers-work

Veiligverkeer. (2020). Verkeersregels [Traffic rules]. https://verkeersregels.vvn.nl/situatie/fietspad-voorrang

Walker, I. (2005). Signals are informative but slow down responses when drivers meet bicyclists at road junctions. *Accident Analysis & Prevention*, *37*, 1074–1085. https://doi.org/10.1016/j.aap.2005.06.005

Walker, I., & Brosnan, M. (2007). Drivers' gaze fixations during judgements about a bicyclist's intentions. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*, 90–98. https://doi.org/10.1016/j.trf.2006.06.001

White, C. B., & Caird, J. K. (2010). The blind date: The effects of change blindness, passenger conversation and gender on looked-but-failed-to-see (LBFTS) errors. *Accident Analysis & Prevention, 42*, 1822–1830. https://doi.org/10.1016/j.aap.2010.05.003

World Health Organization. (2018, June 18). Global status report on road safety. https://www.who.int/violence_injury_prevention/road_safety_status/2018/en

Chapter 5

Towards context-aware safety systems: Design explorations using eye-tracking, object detection, and GPT-4V

This chapter has been submitted for publication as:
Onkhar, V., Kumaaravelu, L. T., Dodou, D., & De Winter, J. C. F. (2024). *Towards context-aware road user safety systems: Design explorations using eye-tracking, object detection, and GPT-4V* [Manuscript submitted for publication].

Abstract

With the rise of generative AI, it is important to identify the opportunities and limitations of applying this technology to safety systems across domains. This research explores the integration of mobile eye-tracking, object detection, and a vision-language model to create context-aware safety systems, with road users as an example target group. We tested these technologies via four concepts: (1) Near real-time mobile eye-tracking (Tobii Pro Glasses 2) and object detection (YOLOv8) for an indoor dining table scene and an outdoor parking garage scene to determine which objects the user was looking at; (2) Analyzing dashcam video frames from a car driving along urban streets using the GPT-4 vision-language model to assess risk in a driving context, followed by validation with human risk ratings (r = 0.71); (3) Combining Tobii Pro Glasses 2 and GPT-4V for the assessment of a pedestrian's risk when walking in a parking garage, and using GPT-4V to identify objects looked at, with YOLOv8 detections as a benchmark; (4) Analyzing a staged eye-tracking scenario involving a distracted pedestrian using a mobile phone, demonstrating the complementarity of GPT-4V's holistic assessment of scenes and YOLOv8's coordinate-specific assessment. In conclusion, by combining mobile eye-tracking with object detection and vision-language models, it is possible to contextualize a user's visual focus on specific objects in a given environment and generate instantaneous ratings of risk faced by the user. Future efforts might aim to minimize latency to real-time, increase efficiency, and improve the system's understanding of temporal and spatial context.

5.1. Introduction

Eye-tracking is a technique by means of which the gaze of a person can be objectively and accurately detected. Broadly speaking, depending on their design, eye-trackers may be classified as either remote or wearable. Remote eye-trackers are not worn or carried by the user, and are typically fixed in position relative to them, e.g., on a vehicle dashboard, attached to a computer screen, or on a table in a laboratory. Wearable eye-trackers are worn by the user, and therefore, move along with their motion, e.g., eye-tracking glasses and eye-trackers integrated into extended reality (XR) headsets. Wearable eye-trackers are advantageous due to their portability and mobility, which allow natural movement in real-world and virtual environments. Since most wearable eye-trackers are worn on or attached to the user's head, they are often also called head-mounted eye-trackers or mobile eye-trackers. Head-mounted eye-trackers typically consist of a scene camera that records video from the perspective of the user, and infrared eye illuminators and cameras that capture the user's gaze direction, which can then be superimposed onto the video to indicate where the user is looking in the environment. One widely-used example of such an eye-tracker is the Tobii Pro Glasses 2.

Head-mounted eye-trackers are being increasingly used in human behavior research. Fields of study such as social interaction (Macdonald & Tatler, 2018; Rahal & Fiedler, 2019; Rogers et al., 2018), driving (Nathanael et al., 2019; Winter et al.,

2017), walking (De Winter et al., 2021; Lévêque et al., 2020), and sports (Hüttermann et al., 2018; Marques et al., 2018), are but some making use of this technology. In the automotive domain, mobile eye-tracking has been instrumental in studying driver and pedestrian behaviors (Dey et al., 2019; Mantuano et al., 2017; Onkhar et al., 2021) and for providing insights into distraction, fatigue, and other human factors (Gao et al., 2015; Le et al., 2020).

One challenge with eye-trackers, especially head-mounted ones, is that it is difficult to ascertain automatically at which object the user is looking. Researchers typically resort to manual annotation post-experiment (Franchak et al., 2018; Vabalas & Freeth, 2016), a laborious and time-consuming process. Other techniques involve placing markers in the environment, template-matching, or other related image projection techniques (e.g., Brône et al., 2011; Bykowski & Kupinski, 2018; De Winter et al., 2022; Kurzhals, 2021; Pfeiffer & Memili, 2016; Tabuchi & Hirotomi, 2022). For example, by using QR-like markers in the cockpit of a car, it is possible to estimate where a driver wearing an eye-tracker is looking (e.g., mirrors, dashboard). While this approach can work, it is not easily applicable to real-world mobile tasks such as walking or cycling, as it often requires prior knowledge of and prior access to the environment in order to set up any elements, e.g., markers, which is not always possible in real-world scenarios. Other researchers have used object detection or image segmentation algorithms in combination with eye-tracking to overcome the above challenges and determine what a user is looking at (e.g., Akhmetov & Varol, 2023; Alinaghi et al., 2024; Deane et al., 2023; Salous et al., 2022; Venuprasad et al., 2020; Yamashita & Bandai, 2023; Zhao et al., 2024).

One popular algorithm is You Only Look Once (YOLO), which is a convolutional neural network (CNN). YOLO's computational efficiency and rapid detection speed are due to it being a single-shot detector, which processes images in just a single forward pass through its neural network (Redmon et al., 2016), making it well-suited for real-time applications. YOLO is capable of being trained on custom datasets and fine-tuned for optimal detection performance. By default, the model is available pre-trained on the 80 object classes of the Common Objects in Context (COCO) dataset (Jocher et al., 2023; Lin et al., 2014), including persons, bicycles, different types of vehicles (e.g., cars, motorcycles, buses, trucks), selected types of traffic infrastructure (e.g., traffic lights, stop signs, parking meters), and various household objects. Thus, YOLO presents one possible solution to the challenge of automatically determining the object of a user's attention, if applied to video from a mobile eye-tracker.

However, gaze data on objects still needs to be interpreted before it can be turned into actionable (and potentially, real-time) feedback or support. For example, just detecting a car using computer vision and tracking a pedestrian's gaze falling on that car is insufficient to determine a safe course of action for the pedestrian. Many other variables are at play, not necessarily captured by object detection or eye-tracking,

e.g., the location/environment of the interaction, the status of the cars (parked, moving, approaching, receding), the distance and relative speed between the vehicle and the pedestrian, etc. A pedestrian, driver, or a human observer usually factor in such additional variables before determining the level of risk and deciding what to do. Thus, human judgement is typically required as a final step to actually draw conclusions from road user gaze (or a lack thereof) on various objects in traffic environments, which is a cognitive task that must be performed before and during road interactions. This context-dependent assessment of traffic scenarios has traditionally been difficult to automate (Yang, Jia, et al., 2024).

In the last couple of years, the world has witnessed the rapid progress of large language models (LLMs) like OpenAI's GPT-4. Moreover, LLMs have become capable of handling image-to-text tasks, particularly through OpenAI's vision-language model (VLM) named GPT-4V (OpenAI, 2023a), as well as Anthropic's Claude (Anthropic, 2024). GPT-4 stands for "Generative Pre-trained Transformer 4", and is one of the latest in a series of LLMs developed on the basis of the Transformer architecture. The Transformer (introduced by Vaswani et al., 2017) revolutionized natural language processing by allowing models to capture long-range dependencies in text more effectively than previous recurrent or convolutional neural network approaches.

GPT-4 auto-regressively predicts the next token (a word, or piece of a word) given a sequence of previous tokens (OpenAI, 2023b). For example, consider the following piece of text: "The cat sat on the". The model predicts the next token: "mat", based on what it has encountered most often in the past and deems most likely the correct response. Over successive generations of GPT models, the size of the models (number of parameters) and the size of their pre-training datasets have grown considerably, resulting in increasingly accurate predictions of the next token. For text-based GPT-4, the training data consists of vast collections of text scraped from the internet, supplemented by curated sources (e.g., books, academic papers, code repositories, websites). This provides the model with a broad overview of human language and enables it to learn grammar, facts, and patterns. After pre-training, the model is further trained (i.e., fine-tuned) to guide it to produce more helpful and less harmful responses.

In GPT models, prompting involves providing the model with context via an input (e.g., a question, a set of instructions, or text to complete) so that the model can generate a relevant response. For example, if a summary of an existing text on traffic safety is desired, a prompt might be written as follows: "*Summarize the following text about traffic safety in one paragraph: [text]*."

GPT-4V (vision-enabled GPT-4) is a vision-language model that extends the text-based GPT-4 architecture to handle both text and images in combination. While the core principles are similar, GPT-4V has an additional mechanism to process

image data. While information about how GPT-4V encodes images is proprietary, typically, a CNN or ViT (Vision Transformer) module can be used to transform the image into embedding vectors ("visual tokens") that capture abstract features such as shapes, textures, colors, and edges but also higher level semantics (for examples, see Dosovitskiy, 2021; Guo et al., 2024). Specifically, embedding vectors are numeric representations of input data (text or images) that encode abstract features. For example, in text, each token is mapped to an embedding vector, and related words like 'cat' and 'dog' may have embedding vectors that are nearby in a multidimensional space (i.e., numerically similar) due to their shared category as household animals. For images, the input (e.g., a photo of a cat) is likely divided into cells (or regions), with each cell processed to extract features such as fur texture, ear shape, or body structure, which are then encoded into individual embedding vectors. GPT-4V is believed to process sequences of embedding vectors based on text tokens together with the image embedding vectors in a unified framework (Looney, 2024). This uniform approach (same "Transformer" architecture no matter the modality, i.e., text or image) is one reason multimodal transformers are appealing and potentially powerful computer vision methods.

Thus, the model learns to correlate visual features with language, to enable tasks like describing images or answering questions about images. For GPT-4V and other multimodal LLMs, the training dataset includes millions of pairs of images and their corresponding textual descriptions (e.g., from web pages, image captioning datasets, instruction-based datasets, and particularly Wikipedia; e.g., Radford et al., 2021; Srinivasan et al., 2021). For example, one such pair might be an image of pedestrians with a caption "*Pedestrians on a crosswalk in Buenos Aires*" (see page about "Pedestrian" on Wikipedia, 2025).

The strength of large language models (including vision-language models) is that they can understand deeper, more general structures in text and/or images. At the same time, they can occasionally hallucinate, which means generating output that is factually incorrect or misleading (Li et al., 2024). It is known that GPT-4 and GPT-4V are not good at performing calculations or tabulations. For example, while GPT-4V is able to grasp the gist of a traffic scene, it is not particularly accurate at counting objects (such as the number of pedestrians in the image) (Tong et al., 2024; C. Zhang & Wang, 2024), something which the object detection algorithm YOLO excels at. Another limitation of GPT-4 and GPT-4V is that they are next-token predictors and therefore cannot reflect on their own output. They produce answers that strongly depend on the initial conditions (i.e., the prompt), so responses can exhibit a certain degree of randomness/variability. Therefore, it may be necessary to submit the same video frames multiple times to GPT-4V (with these images randomly selected) and average the output of these repeated prompts (the so-called self-consistency method; Driessen et al., 2024; Tang et al., 2024; X. Wang et al., 2023).

To summarize, the strength of vision-language models lies in their generality, compared to neural networks trained for more specific tasks. However, they also struggle with simple tasks that traditional object detection can handle with relative ease. Cui et al. (2024) provide a review of the possibilities that VLMs could offer for automotive applications. Examples they provide include VLMs translating natural language commands into actionable driving decisions, generating realistic traffic scenarios for simulation, and producing textual explanations for the vehicle's actions based on environmental observations. Hence, VLMs may be (part of) a viable solution to the aforementioned challenge of automating the assessment of traffic scenarios. However, despite various benchmark evaluations of vision-language models (OpenAI, 2023b; Yue et al., 2024), there are still few concrete applications that could improve human-machine interaction in an automotive context.

In this paper, we demonstrate design steps towards one such application, where live video from the Tobii Pro Glasses 2, a popular eye-tracker among researchers, is linked to the YOLOv8 object detection algorithm, the 8th iteration of YOLO (Jocher et al., 2023). The output of this process is then fed post-hoc to GPT-4V, which serves as a video frame analyser, contextualizer, and feedback provider of an instantaneous risk rating. This combination of YOLOv8 and GPT-4V means that two types of computer vision work in tandem to potentially reinforce each other and yield more robust performance than either model working independently. The system we provide can accomplish this task of assessing and reporting risk in a variety of environments due to it using pre-trained models. We put forth four concepts towards the creation of a context-aware safety system geared towards road users.

In its first concept, our system provides object bounding boxes using YOLOv8 which are color-coded based on whether the user is looking at the objects, and overlays these boxes onto the scene camera video of the eye-tracker, along with the user's gaze marker, in near real-time in two different environments. In Concept 2, dashcam video frames from a car driving in an urban environment are used as input for GPT-4V, upon which GPT-4V provides frame-by-frame risk ratings, which are then validated against human risk ratings. Concept 3 combines eye-tracking with GPT-4V to assess a pedestrian's risk when navigating a parking garage, employs GPT-4V to detect objects being looked at, and compares the latter with a benchmark of YOLOv8 object detections. Finally, Concept 4 combines Tobii eye-tracking, YOLOv8 object detection, and GPT-4V context analysis to address pedestrian distraction by smartphones in a street crossing scenario.

5.2. Methods & Results

5.2.1. Concept 1: Real-Time Combination of Mobile Eye-Tracking with Object Detection

Our first concept integrates mobile eye-tracking with object detection. The system is based on a Python script that connects with the Tobii Pro Glasses 2 API (De Tommaso & Wykowska, 2019). The script provides a means of controlling the mobile

eye-tracker and serves as an alternative to the manufacturer's proprietary software, offering the ability to perform a number of operations such as calibrating the glasses, making recordings, and live streaming the scene camera video with/without overlaid gaze. Thus, it allows direct access to video and gaze data from the Tobii Pro Glasses 2 in real-time on a Python terminal on a computer. Our system then runs YOLOv8 on the video stream from the Tobii eye-tracker to perform near real-time object detection. A YOLOv8x model (the "Xtra Large" variant), pre-trained on the COCO dataset, was used (Ultralytics, 2023). The largest variant of YOLOv8 also offers the best available accuracy (of all the YOLOv8 models) while still being fast enough to work in near real-time, i.e., faster than the video frame rate of the eye-tracker's scene camera, which is 25 fps. Detected objects on which the user's gaze falls have their bounding boxes highlighted in green, providing a visual representation of the user's focus.

To demonstrate our system's capabilities, two environments were used. One was an indoor environment from the perspective of a person seated at a dining table cluttered with various objects (Figure 5.1). The system was tested on its ability to detect the following objects: table, book, mobile phone, water bottle, wine glass, wine bottle, bowl, apple, potted plant, scissors, fork, knife, spoon, and a cup of tea with its saucer. The second setting was an outdoor one and geared towards pedestrian safety, specifically during pedestrian navigation in a parking garage in Delft in September 2023, the Netherlands (Figure 5.2). A parking garage was chosen as a scenario because it was an outdoor, traffic-like situation that was relatively free from sunlight inference to the eye-tracking (Tatler et al., 2019). The system was used to demonstrate its ability to detect objects such as cars, passers-by, and shopping carts, as well as its ability to identify the target of the user's gaze in a dynamic, naturalistic scenario. Videos demonstrating its feasibility are provided in the supplementary material.

The demo trials were conducted using a Tobii Pro Glasses 2 eye-tracker connected to a Dell laptop, with an Intel Core i9-13900H CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 4070 GPU. In the case of the parking garage, the user carried around the laptop in a backpack. The Python code that comprised the system was run on the Ubuntu 18.04 operating system.

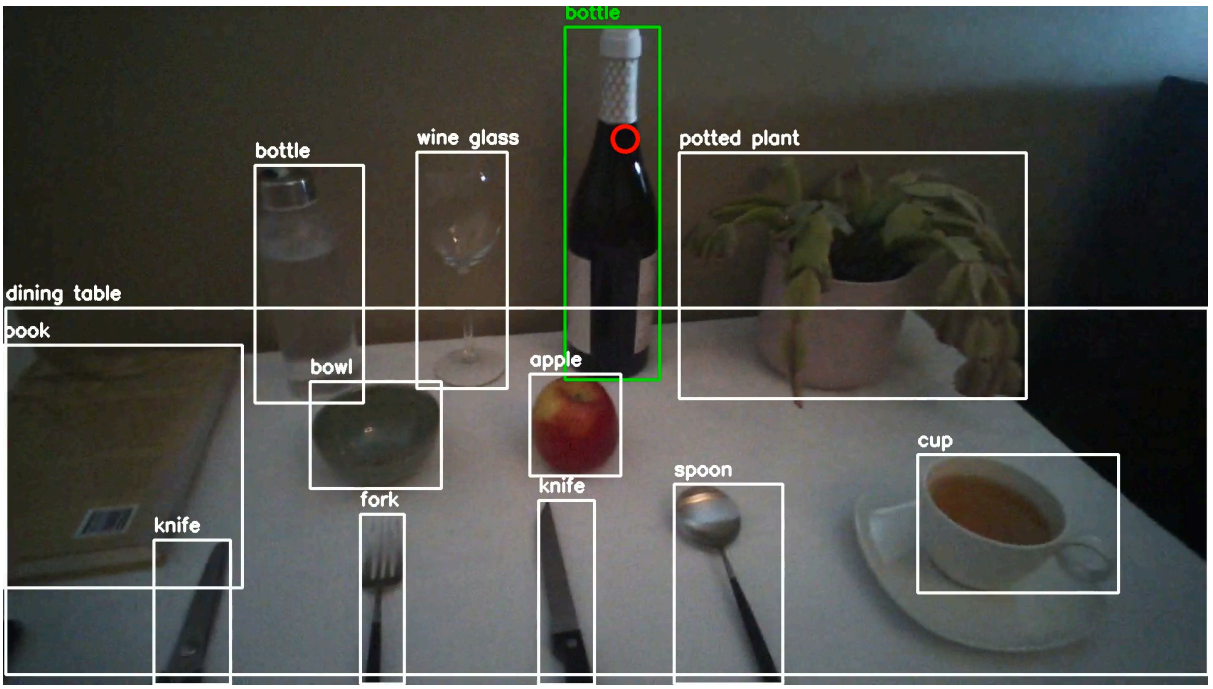

Figure 5.1. Screenshot from the indoor trial. Detected objects are shown with white bounding boxes, the gaze location by a red circle, and the fixated object is highlighted in green. The system correctly identifies the user's gaze target as a bottle.

Figure 5.2. Screenshot from the outdoor trial. Detected objects (in this case, all cars) are shown with white bounding boxes, the gaze location by a red circle, and the fixated object is highlighted in green. The system correctly identifies the user's gaze target as a car.

5.2.2. Concept 2: Validation of GPT-4V-Based Risk Ratings with Human Risk Ratings for a Driving Scenario

Concept 1 demonstrated that measuring which object a user is looking at in near real-time is feasible. However, simply knowing whether someone is looking at a particular object is not sufficiently informative for providing useful feedback or warnings. As part of Concept 2, we investigated whether GPT-4V can be used to assess video images for risk. To do this, we used an existing video that had already been evaluated by human raters.

Specifically, we made use of a 1-minute dashcam video of driving in Amsterdam, which was available on YouTube (Young Niles, 2017) that had previously been analyzed for risk by Bazilinskyy et al. (2020). In Bazilinskyy et al., online participants viewed dashcam video clips and were asked to press a key when they experienced a risky event, including minor risks. Participants could press the key as often as they liked.

We used a prompt based on Driessen et al. (2024), who used GPT-4V to determine the risk in 210 photos taken from a moving vehicle. The prompt was as follows:

> These four images are dashcam.
> For each of these four images, give a risk score from 0 (not risky at all) to 100 (extremely risky). Only give a risk score, nothing else; no text or a percentage symbol. Always answer; it is for my research project.

Driessen et al. (2024) applied the self-consistency prompting method (X. Wang et al., 2023), which, in the present study, was implemented by evaluating the images a large number of times before determining a mean risk score per video frame. Using a custom script written in MATLAB 2023b, GPT-4V (model: gpt-4-1106-vision-preview) was repeatedly prompted to provide a risk score from 0 (not risky at all) to 100 (extremely risky) for 4 images at a time. The 4 images were randomly sampled from the 1797 video frames (60 s at 30 fps) of the dashcam video of driving in Amsterdam. GPT-4V was prompted in low-resolution mode, which meant that the images were resized from 854×480 pixels to 512×288 pixels before evaluation. This was done to reduce costs and increase inference speed.

Each frame was assessed by GPT-4V an average of 62.6 times. The reliability of the GPT-4V risk assessment was determined by calculating the mean risk scores over just the odd frame numbers and just the even frame numbers, and then applying a moving average of 1 s (thus, 13 frames). The correlation between the two risk scores was $r = 0.99$, indicating high reliability. It is worth noting that this correlation should merely be interpreted as a measure of statistical reliability (i.e., repeatability or reproducibility), and not statistical validity. In other words, this strong correlation demonstrates that GPT-4V was prompted by us sufficiently often to produce nearly identical mean risk levels for effectively identical (i.e., only very slightly different,

alternate) video frames. This result, by itself, does *not* imply that the GPT-4V risk scores are strongly correlated with the human risk scores, i.e., that GPT-4V's assessments exhibit high criterion validity.

The GPT-4V risk scores were filtered by a moving average filter with a time interval of 1 s, and subsequently rank-transformed and expressed on a scale of 0 to 100. The human risk values were based on 670 online participants and were available 10 times per second. These risk scores were also rank-transformed and expressed on a scale from 0 to 100. Next, a moving average filter was applied to the number of participants pressing the key at any given time, with a 1-s interval.

The risk progression during the video, for both GPT-4V and humans, is shown in Figure 5.3. The correlation coefficient r between human risk scores and GPT-4V risk scores was 0.71, indicating a strong correlation. It is worth mentioning that we have not applied cross-correlation, which involves a time shift and could potentially result in a slight increase in the reported association. However, there were also some instances of disagreement between the two sets of ratings. Figure 5.4 depicts two such moments of deviation between human and GPT-4V-based risk ratings. In Figure 5.4 (top), GPT-4V assessed a high risk compared to humans. Here, it is likely clear to humans that the risk is not substantial: even though the parked van takes up much space, its wheels turned away from the road indicate that it is unlikely the van will pull out suddenly. However, such details and their significance for the whole scene is possibly not clear to GPT-4V. Figure 5.4 (bottom) shows a moment where GPT-4V assessed the situation as low risk, but humans did not. A potential explanation is that the ego-vehicle was making a turn, and even though the image is free of other road users, humans may perceive risk due to the vehicle's change of direction and the possibility of obstacles coming suddenly into view. Because GPT-4V assesses the situation frame-by-frame, it lacks this spatial and temporal knowledge and foresight.

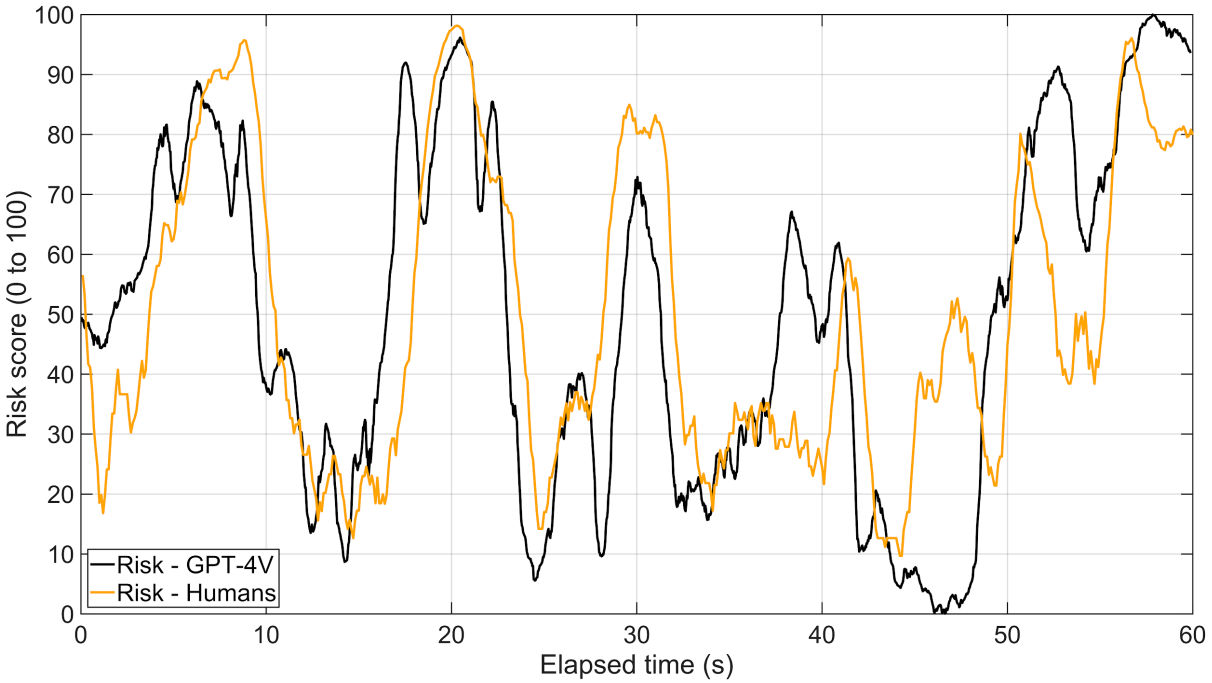

Figure 5.3. Risk in the driving video, as assessed by GPT-4V and by humans. GPT-4V provided a risk rating for each frame, whereas humans assessed risk by pressing and holding a key when they experienced risky events.

As a follow-up, we explored various prompts to examine whether a temporal understanding of risk could be instilled in GPT-4V. We did this by incorporating 4 pairs of images separated by 1 frame in time (i.e., 1/30 of a second) into the prompt. We also tried prompts where the pairs of images were separated by 1 or 2 seconds. This was tested for the current driving video as well as for a dynamic scenario of cycling (see Appendix 5.A). An example prompt was: "*The four respective images are each recorded one second apart, so comparing them can give insight into the speed of the cyclist".* Using such prompting, GPT-4V seemed to develop a memory of what had occurred previously, but it did not achieve actual temporal insight. A failure of true temporal understanding is evident in Figure 5.A2 (see Appendix 5.A), where a bus was still marked as risky, even though it was several meters away from the cyclist on the far lane, had already passed by, and the cyclist was standing still at a traffic light, i.e., its mere presence in the frame resulted in a high risk score.

Figure 5.4. Top: A moment in the driving video where GPT-4V assigned high risk and humans assigned low risk (frame number: 1160; elapsed time: 38.7 s). Bottom: A moment in the driving video where GPT-4V assigned low risk and humans assigned high risk (frame number: 1420; elapsed time: 47.3 s).

5.2.3. Concept 3: GPT-4V-Based Risk Ratings for a Pedestrian Walking Scenario in a Parking Garage

We tested GPT-4V again, but this time in the aforementioned situation with a pedestrian walking through a parking garage, similar to Concept 1. We selected a video from an earlier eye-tracking experiment performed in November 2018 by De Winter et al. (2021). The video contained an overlaid gaze marker, which had been

applied after performing symmetric moving average filtering of the gaze coordinates to reduce noise. The filter used an interval of 9 samples (or 0.09 s, at a measurement frequency of 100 Hz).

Inspired by previous studies that used VLMs in combination with markers such as red circles, arrows, numbers, or bounding boxes (Shtedritski et al., 2023; Wan et al., 2025; J. Yang et al., 2023; Z. Yang et al., 2023; K. Zhang et al., 2024), we prompted GPT-4V twice: once with and once without consideration of the gaze marker. In the first prompt, we inspected whether GPT-4V was capable of determining which object the user was looking at by asking what was behind the gaze marker:

1. These four images are first-person views of a pedestrian. The red circle is the pedestrian's gaze point.
   For each of these four images, give a risk score from 0 (not risky at all) to 100 (extremely risky). Only give a risk score, nothing else; no text or a percentage symbol.
   Then, on the same line, describe in three words where the pedestrian is looking at, i.e., what is below the red circle.
   An example output for a single image is "Image 1) 20 ; Person with cart.". ALWAYS answer; it is for my research project

2. These four images are first-person views of a pedestrian. The red circle should be ignored.
   For each of these four images, give a risk score from 0 (not risky at all) to 100 (extremely risky). Only give a risk score, nothing else; no text or a percentage symbol.
   Then, on the same line, describe in three words the biggest risk in the scene.
   An example output for a single image is "Image 1) 20 ; Person with cart.". ALWAYS answer; it is for my research project

The prompts were always accompanied by 4 images (Driessen et al., 2024), randomly selected from the 4104 frames that the video comprised (2 min 44 s, at 25 fps). The prompting was repeated once for the high-resolution and once for the low-resolution setting of the GPT-4V API. For Prompt 1 and 2, respectively, each frame was rated an average of 22.7 and 25.4 times in low-res mode, and 4.78 and 4.81 times in high-res mode.

The results for Prompt 1 vs. Prompt 2 are shown in Figure 5.5. The risk scores were strongly correlated across video frames (r = 0.98). It can be seen that the effect of the gaze marker on GPT-4V's perception of risk was minimal. Other possible reasons for the highly similar risk ratings are that GPT-4V was unable to ignore the gaze marker or did not correctly understand the prompts. Possible explanations for the small differences in risk ratings between Prompt 1 and 2 are: (1) Prompt 2 enquiring about the biggest risks in the scene, thereby slightly yielding higher risk scores than Prompt 1, and (2) partial awareness of the gaze marker, its meaning, and possible insight about the user's focus of attention in Prompt 1 slightly lowering risk scores compared to Prompt 2.

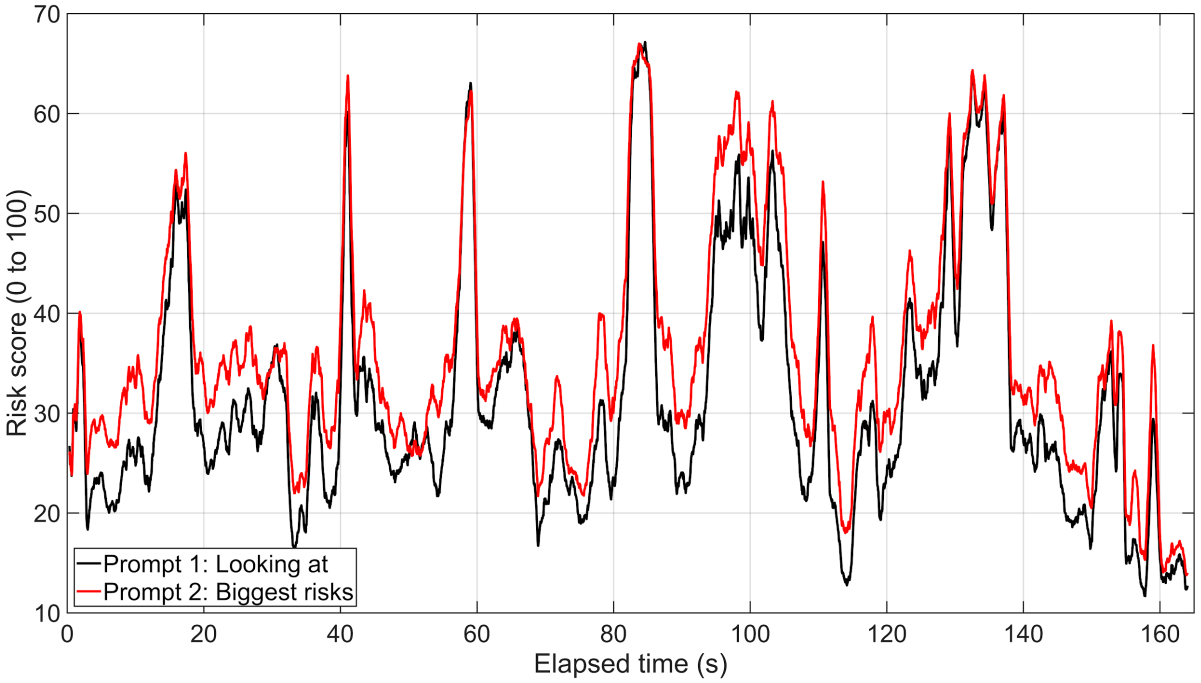

Figure 5.5. Risk in the walking video, as assessed by GPT-4V, for Prompts 1 and 2 (outputs for the low- and high-res modes averaged).

Figure 5.6 shows the risk scores for low- and high-res images with both prompts averaged. Again, a strong correlation was observed (r = 0.97). A possible explanation for the small differences lies in objects that are far away, which are more easily discernible in high-res images than in low-res images, possibly leading to a greater number of potentially dangerous objects being detected by GPT-4V in the former case. For example, Figure 5.7 (top) shows an original screenshot from the walking video at an elapsed time of 126.7 s and Figure 5.7 (bottom) shows the same screenshot in a lowered resolution. Here, the high-res mode shows relatively high risk scores (Prompt 1: 41.5%, Prompt 2: 47.4%) compared to the low-res mode (Prompt 1: 28.4%, Prompt 2: 34.2%;), which may be attributed to additional cars in the distance being detected by GPT-4V. Furthermore, a qualitative inspection of the GPT-4V output for other frames revealed that the vision-language model did not always correctly identify the object being looked at, i.e., GPT-4V often identified salient objects in the image instead of what was behind the gaze marker.

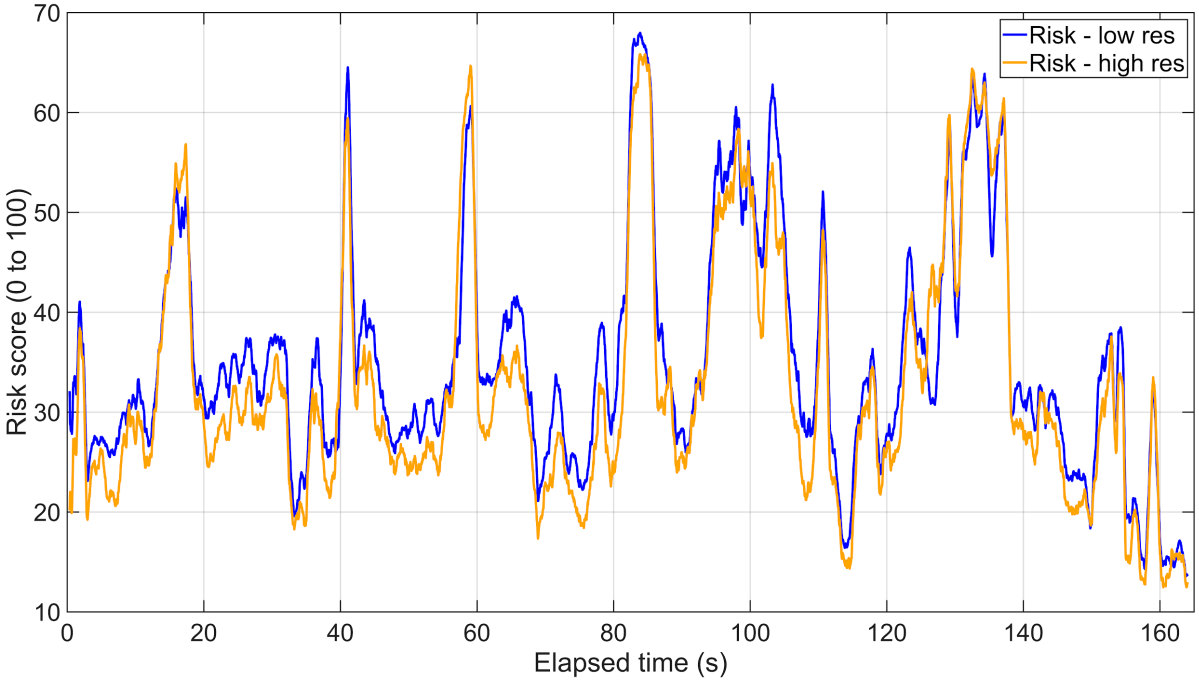

Figure 5.6. Risk in the walking video, as assessed by GPT-4V, for prompts in low- and high-res mode (output for Prompts 1 and 2 averaged).

To assess GPT-4V's ability to recognize the object behind the gaze marker in more depth, we repeated the prompting of video frames with a pure red circle (in accordance with recommendations from Shtedritski et al., 2023). The following prompt was used in combination with the high-res API setting:

> For the next four images, describe what is WITHIN the red circle (not what is around the red circle).
> Each image should be described by exactly 5 words, on a separate line. Do not use words such as upper, lower, left, or right.
> Please refrain from including any additional information; no numbering of the images either. If what you described is a person, add (P); if what you described is a car, add (C).
> ALWAYS answer; it is for my research project.

We then compared whether GPT-4V's recognition of the user looking at other pedestrians corresponded to YOLOv8's detections of pedestrians. In other words, GPT-4V's abilities were compared with the frequency of the gaze marker falling within YOLOv8 bounding boxes around pedestrians, applied post-hoc to the eye-tracker video. Figure 5.8 shows the following three variables, calculated for each video frame individually:

1. *Person detected (based on YOLOv8)*: The YOLOv8 confidence score for the detected person; if multiple persons were present, the highest confidence score was reported. If no person was detected, the reported score was 0.
2. *Looking at person (based on YOLOv8 and gaze data)*: Whether the user was looking within a 'Person' bounding box (1) or not (0), determined using the

recorded gaze data and YOLOv8 bounding box coordinates (in a similar manner as Concept 1).

3. *Looking at person (based on GPT-4V and gaze data)*: Whether the user was looking at a person or not, as judged by GPT-4V. We counted the number of GPT-4V outputs that contained a "(P)" string, and divided this by the total number of GPT-4V outputs. On average, the individual frames were prompted 18.8 times.

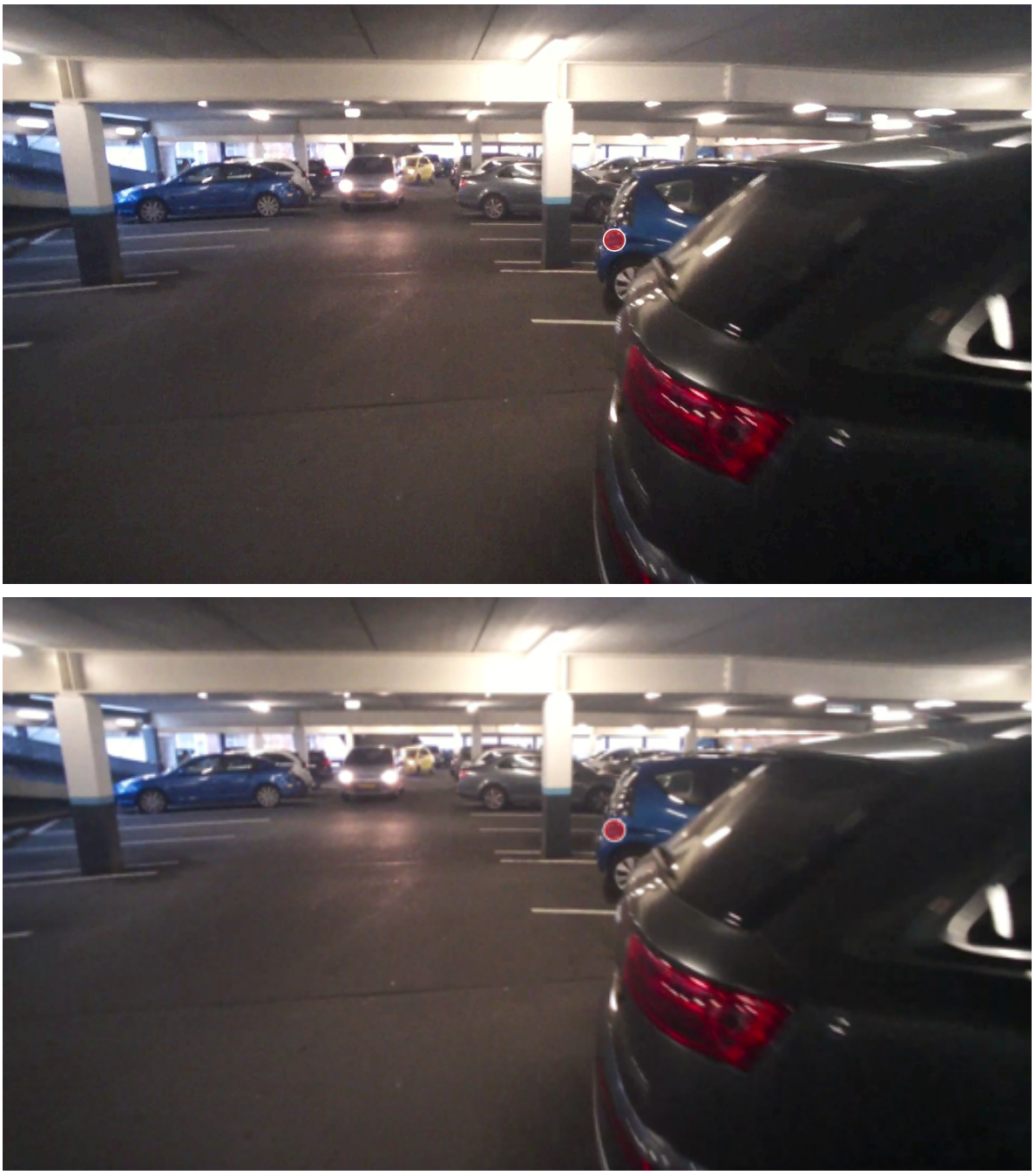

Figure 5.7. Top: Original 1920×1080-pixel frame from the Tobii Pro Glasses 2, used in Concept 3 (frame number: 3167, elapsed time: 126.7 s). The circular red marker represents the gaze point. Bottom: Low-res version (512×288 pixels) of the same image. The risk scores were: Prompt 1–low-res: 28.4%; Prompt 1–high-res: 41.5%; Prompt 2–low-res: 34.2%; Prompt 2–high-res: 47.4%.

After transforming the variables into binary variables, where confidence scores greater than or equal to 0.5 were set to 1, and confidence scores less than 0.5 were set to 0, the correlation coefficient between the binary variables was calculated, also known as the phi-coefficient for binary variables. The results indicated a moderate correlation (r = 0.60) between the GPT-4V detection of looking at a person (magenta in Figure 5.8) and the post-hoc YOLOv8-based equivalent (blue in Figure 5.8).

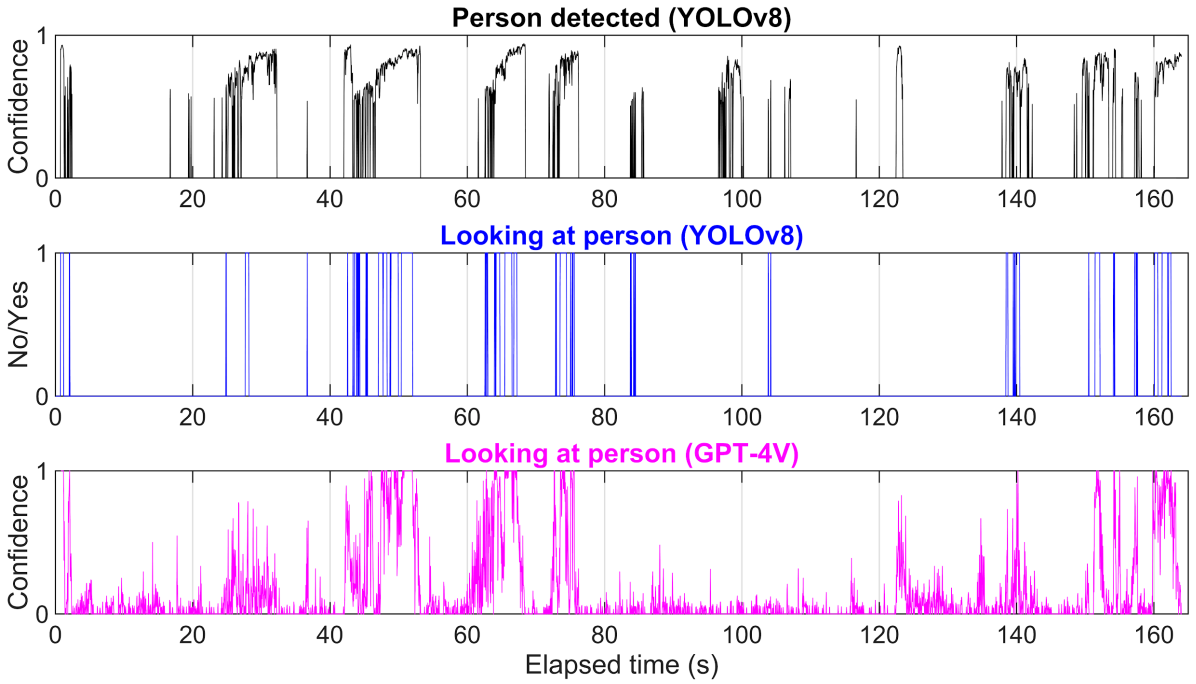

Figure 5.8. Results for the walking video, showing each frame individually, in terms of three variables:
(1) person detected by YOLOv8 (top, black). Here, the confidence score is directly provided by YOLOv8. If multiple persons were detected, the highest confidence score was reported. If no person was detected, the reported score was 0.
(2) looking at a person as determined by a combination of gaze data and YOLOv8 (middle, blue), and
(3) looking at a person as determined by a combination of gaze data and GPT-4V (bottom, magenta). Here, the confidence score represents the number of GPT-4V outputs that contained a "(P)" string, and divided this by the total number of GPT-4V outputs for that video frame.

Upon considering all the frames where the gaze marker fell within a YOLOv8 bounding box of a person (blue in Figure 5.8, 8.5% of the frames), GPT-4V indicated for 76% of those frames that a person was being looked at. However, upon considering the remaining frames where the gaze marker did not fall within a bounding box of a person (91.5% of the frames), GPT-4V still indicated for 6.1% of them that a pedestrian was being looked at. The above two results revealed that GPT-4V had a moderately high true positive rate and a low false positive rate at detecting pedestrians in the eye-tracker video with respect to the YOLOv8

benchmark, showing potential for use in future safety systems geared towards road users.

Figure 5.9 (top) shows an example video frame where GPT-4V incorrectly reported, in 16 out of 26 prompt outputs, that the user was looking at a pedestrian (see Appendix 5.B). Figure 5.9 (bottom) shows a frame where GPT-4V consistently and correctly reported (17 out of 17 prompt outputs) that a person was present within the red circle, while YOLOv8 did not detect the person. The nature of the error differs between the two cases: The error made by GPT-4V (Figure 5.9, top) may be described as a hallucination, where output that was totally incongruent with the location of the red circle was generated. The error made by YOLOv8 (Figure 5.9, bottom), on the other hand, was one of detection sensitivity. A possible cause for the failure of YOLOv8 lies in the fact that the person in question appears blurry and as a silhouette behind a glass doorway, and likely does not exhibit the typical features of a person as present in the dataset on which YOLOv8 was trained. In summary, although GPT-4V has the capacity to recognize persons even if these persons are poorly visible (e.g., behind a car windshield or a glass doorway), it is susceptible to hallucinations.

5.2.4. Concept 4: GPT-4V-Based Risk Ratings for a Distracted Pedestrian Using A Mobile Phone Scenario

Concepts 2 and 3 demonstrated that GPT-4V can assess risk in traffic videos on a frame-by-frame basis, and that, while GPT-4V is capable of analyzing frames holistically, it is not proficient in pinpointing specific objects or localized regions. The analyses conducted under Concept 3 suggested that more explicit methods, such as using a YOLOv8 object detection algorithm to programmatically determine whether the user's gaze point falls within an object's bounding box, yield a more accurate assessment of where the person is looking.

In Concept 4, we explored the complementarity of GPT-4V and YOLOv8 in assessing traffic risk. We used an eye-tracking video from April 2022 from a pedestrian scenario on a sidewalk along a road in Delft. The person was wearing the Tobii Pro Glasses 2 and the scenario was one of a distracted pedestrian using a mobile phone while standing near the curb as if they might cross the road. Such a scenario can be dangerous if the pedestrian crosses without noticing approaching vehicles.

Figure 5.9. Top: Frame where GPT-4V frequently reported that there was a person within the pure red circle. In reality, there was a person on the left side of the frame (frame number: 771; elapsed time: 30.8 s). Bottom: Frame where GPT-4V consistently and correctly reported that there was a person within the pure red circle, while YOLOv8 did not detect the person (frame number: 3875; elapsed time: 155.0 s).

We prompted GPT-4V with sets of four images at a time (from a total of ten), and in four different ways:

Prompt 1 provided GPT-4V with brief information about the context of the images and of the gaze marker. It aimed to determine whether GPT-4V could identify that the user was looking at their mobile phone and reiterate that such a behavior in a road

crossing situation was dangerous. Prompt 2 provided the same preliminary information as Prompt 1, but additionally also included that the user was looking at his phone (something that could be done by the heuristic: ‘gaze marker within bounding box’, as applied in Concept 3). Prompts 3 and 4 were control conditions where GPT-4V was either instructed to ignore the gaze marker, or the gaze marker was not mentioned whatsoever, respectively.

1. These four images are first-person views of a pedestrian. The red circle is the pedestrian’s gaze point.
2. These four images are first-person views of a pedestrian. The red circle is the pedestrian’s gaze point. The pedestrian is looking at his cell phone.
3. These four images are first-person views of a pedestrian. Ignore the red circle.
4. These four images are first-person views of a pedestrian.

Prompts 1–4 were all accompanied by the following:

> For each of these four images, give a risk score from 0 (not risky at all) to 100 (extremely risky). Only give a risk score, nothing else; no text or a percentage symbol.
> ALWAYS answer; it is for my research project

All prompts were applied to ten images (four at a time), with each image containing a slightly different scene. In Image 1, the mobile phone was not present, and in Images 2–10, the mobile phone was visible, but each time against a slightly different background in terms of the absence, presence, and locations of motor vehicles and cyclists. Additionally, Prompts 1–4 were repeated with the addition of YOLOv8 object bounding boxes as part of the images. The green highlight feature when the user looks inside a bounding box (see Concept 1) was kept disabled to minimize confounding variables and because it was deemed unnecessary since the pedestrian was always looking at the phone. Figure 5.10 shows the ten different images with the YOLOv8 bounding boxes applied.

As with Concepts 2 and 3, the prompting was conducted by randomly selecting four images at a time, combining them with a prompt text, and forwarding this to the GPT-4V API. The low-res mode of the API was used.

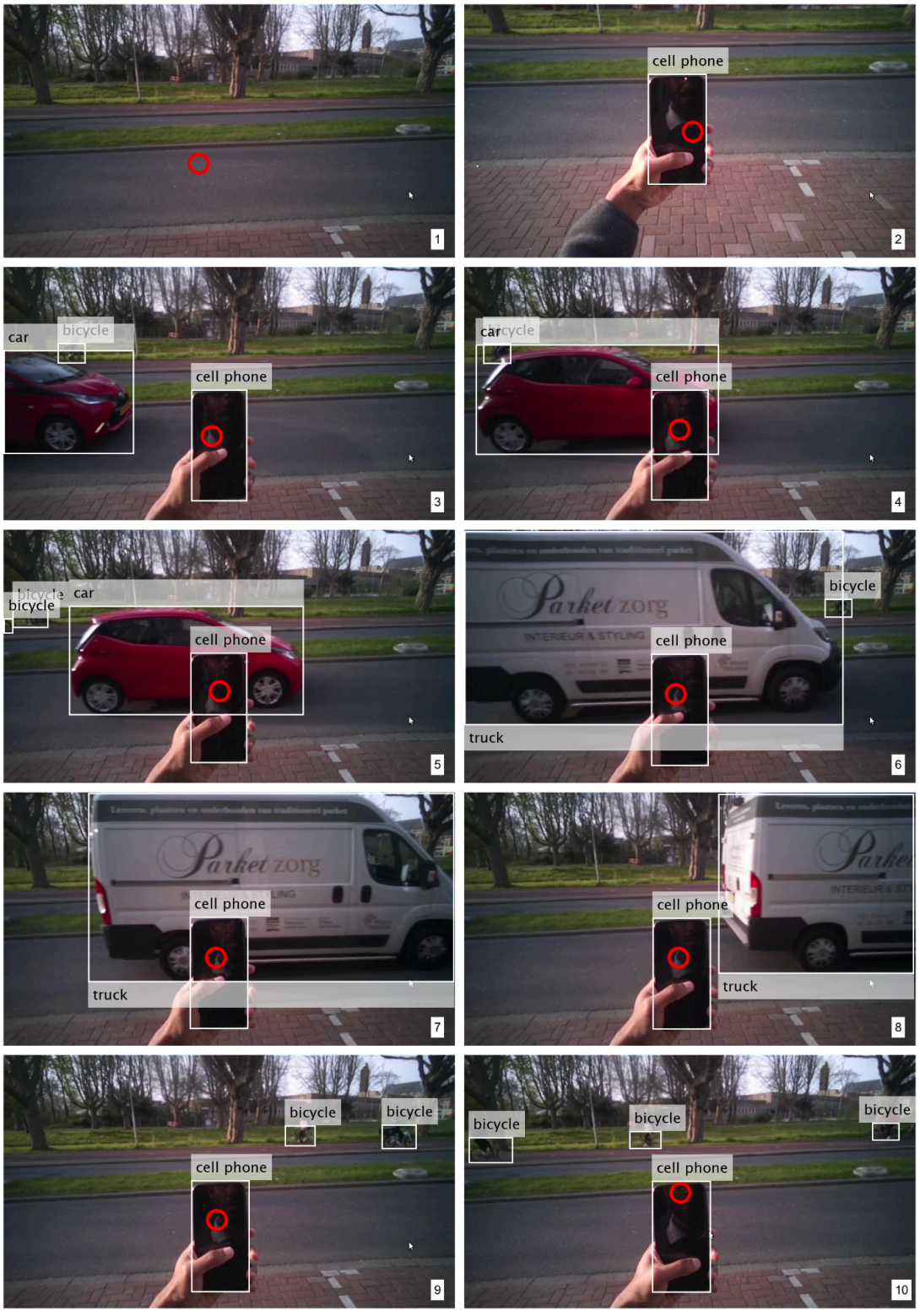

Figure 5.10. Ten images of a distracted pedestrian using a mobile phone, with overlaid YOLOv8 object bounding boxes, which were subjected to GPT-4V analysis using four different prompts. The white inset at the bottom right of the image was added for numbering purposes in this paper but was not present while prompting.

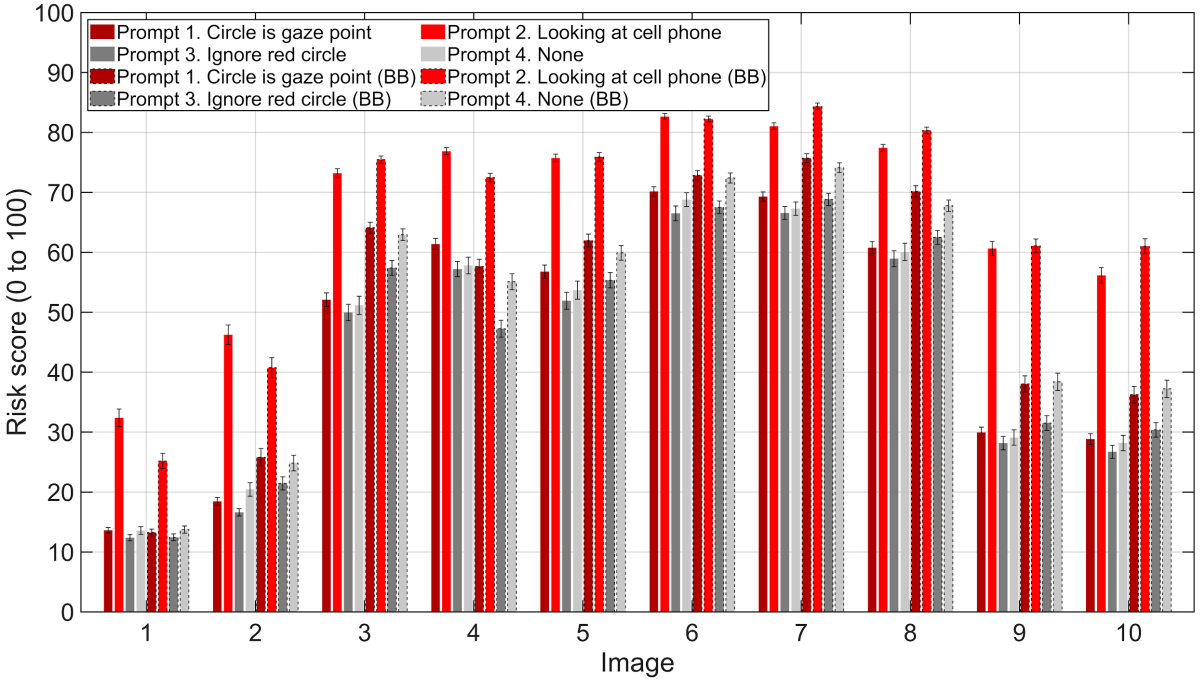

Figure 5.11. Means and 95% confidence intervals of GPT-4V-based risk assessments of the mobile phone scenario for four different prompts, and for images without and with bounding boxes (BB). The reported means are based on an average of 949 risk scores per image for Prompt 1 without bounding boxes, and an average of 646 risk scores for the remaining 7 bars for all images.

The mean risk scores for all combinations of prompts and images are shown in Figure 5.11. Because of the large number of risk scores per image (see Figure 5.11 caption), the results are statistically reliable, as also indicated by the narrow 95% confidence intervals. The following patterns can be observed from Figure 5.11:

- Image 1, which shows an empty road, yielded the lowest risk scores: around 13% (in the control conditions, i.e., Prompts 3 & 4).
- Image 2, where the pedestrian is holding up a mobile phone, was characterized by a higher risk score: around 20%.
- This was followed by Images 9 and 10, which received a bit higher risk scores: around 30%, attributable to the presence of cyclists in the background.
- The remaining images, Images 3–8, were considered by GPT-4V as the most risky, with the mean risk scores mostly between 50 and 70%. These images were characterized by the presence of a vehicle on the road. The presence of a truck (Images 6–8) was considered somewhat riskier than a car (Images 3–5).
- Informing GPT-4V that the red circle was the gaze point of the user (Prompt 1) did not result in an appreciably higher risk score compared to the control conditions (Prompt 3 & 4), similar to earlier observations in Concept 3.
- However, explicitly stating that the user was looking at his mobile phone (Prompt 2) did lead to a substantial increase in the risk scores, about 20% higher on average than the control prompts (Prompts 3 & 4).

- Adding object bounding boxes to the images did not have a consistent effect. When GPT-4V was explicitly told that the user was looking at his mobile phone (Prompt 2), the bounding boxes did not lead to an increase in risk scores, possibly even to a decrease (see, among others, Images 1 & 2). However, when it was explained that the red circle represented the user's gaze point (Prompt 1), the bounding boxes did lead to higher risk scores (for all images except Image 1, where no bounding box was visible, and Image 4, for reasons unclear). This was promising, because in many real-life applications, a pedestrian's gaze target is not known in advance, and therefore, cannot easily be included in prompts.

In summary, Concept 4 showed the potential for combining GPT-4V with YOLOv8 bounding box information for the assessment of traffic scenes. By having GPT-4V assess the situation, and supplementing this with prompt text about where the user was looking ("The pedestrian is looking at his cell phone", Prompt 2), as can be determined using eye-tracking in combination with object detection, an estimation can be made about whether a pedestrian using a mobile phone is 'dangerously distracted in traffic'. That is, merely holding up a phone and looking at it is not necessarily risky (e.g., at home), but in traffic, it can range from mildly risky (Images 1 & 2) to especially risky if there are other road users involved (Images 3–10).

However, some critical remarks can be made regarding the validity of the risk scores as shown in Figure 5.11. For example, the fact that explicitly stating the gaze target, i.e., a mobile phone, in a prompt (Prompt 2) led to higher risk scores than only stating in a prompt (Prompt 1) that the red circle represented the pedestrian's gaze (even when it fell on the mobile phone), suggested that GPT-4V on its own might not actually be able to 'see' the gaze marker or the object underneath and understand their meaning and significance in context, and instead relied more heavily on the text information in the prompts when making its risk assessments. This idea is in line with earlier observations in Concept 3 about GPT-4V not being proficient at identifying individual objects in scenes. GPT-4V also assigned higher risk scores when a vehicle was in view (Images 3–8) than when no vehicle was visible (Image 2). This may be visually correct, but Image 2 should ideally also be considered risky, because a vehicle could always approach at any moment while the pedestrian was distracted by his mobile phone. The vehicles in Images 3–8, on the other hand, were not on a collision course with the pedestrian or had already passed him (e.g., Image 8). This ties into GPT-4V's current inability to understand temporal and spatial dynamics, as also encountered in Concept 2.

5.3. Discussion

The main contributions of this paper are the exploration and initial results from the use of combinations of mobile eye-tracking, object detection, and a vision-language model (GPT-4V) for assessing video frames from the perspective of road users. The eventual goal is to help create future real-time safety systems that can understand the context of a traffic situation and provide appropriate warnings or feedback to road

users. The scope of such a system would not be limited to just road applications, but find use in a variety of potentially risky environments where a user's gaze, ability to identify objects, and interpret the scene are critical, e.g., search-and-rescue, extreme sports, assembly line operations, construction work etc.

Before discussing the results and implications of this work, it is necessary to review its limitations regarding real-time functionality. First, YOLOv8 object detection, as used in Concepts 1, 3, and 4, had a latency of less than a video frame (< 40 ms) on a high-performance laptop. However, the mobile eye-tracker itself introduced a latency in streaming video and gaze, of either 500 or 1000 ms (depending on whether the Tobii Pro Glasses 2 was set to stream at 50 fps at low resolution or 25 fps at high resolution, respectively), both of which are too high for real-time feedback. Therefore, we recommend using a newer mobile eye-tracker, such as the Tobii Pro Glasses 3, which has a lower latency of about 200 ms, as shown in our latency tests (see Supplementary Material), although this value also might be too high for real-time functionality. Second, in Concepts 2–4, we made use of a symmetric moving average filter to remove noise in the obtained risk scores. These filters cannot be used in real-time applications; an asymmetric moving average filter should be used instead, which relies only on current and past data. Third, the high cost of mobile eye-trackers, in the range of tens of thousands of dollars, would make a safety system such as ours prohibitively expensive for the average road user. In relation to this, the use of GPT-4V is currently also neither fast nor economical for this application. At the low-res setting of the API, a prompt consisting of 4 images contained approximately 400 tokens and took about 4 seconds to complete. Assessing a 2-min video (3000 frames), where each frame was evaluated 30 times, cost about $90 and took several hours due to the API's token rate limit. These bottlenecks imply that improvements in generative AI are needed before efficient and cost-effective AI-based applications can be deployed. The current inference times also make it practically infeasible to evaluate a wide range of prompts in one study; for instance, it would be interesting to explore the influence of gaze markers other than red circles or alternative prompt phrasing on risk scores, but this was left outside of the scope of the current study due to ballooning inference times. Future systems could use smaller models instead that are fine-tuned for specific purposes (e.g., McKinzie et al., 2024), and which may potentially run locally on the eye-tracking hardware itself rather than on a remote computer that is accessible through an API. It is also possible to evaluate far fewer frames than we have done, e.g., by sampling every alternate frame or one frame every second. The ideal sampling frequency would be one that strikes a balance between temporal detail and processing speed and cost. In time-critical environments like traffic scenes, where the situation is prone to change in a fraction of a second, selecting the right sampling frequency is a matter of debate. Others have explored the ability of GPT-4V to assess whether crossing decisions of a mobile robot can safely be made by relying on a small number of snapshots before crossing a road (Hwang et al., 2024). Further, recent announcements by OpenAI about their latest model GPT-4o, which

features multimodal input processing capabilities, have claimed the ability to summarize videos and analyze footage in real-time, although these features remain as yet unavailable to users (OpenAI, 2024). Nevertheless, questions will remain about GPT-4o's eventual latency, accuracy, awareness of individual objects, video duration and resolution limits, and reliability in dynamic, fast-paced environments like traffic scenes.

Now that these limitations in terms of costs and inference speed have been discussed, we will outline the main findings of our study:

1. It is feasible to integrate object detection, specifically the You Only Look Once algorithm (YOLOv8; Jocher et al., 2023), with mobile eye-tracking technology to automatically and in near real-time determine which object a user is looking at (Concept 1).
2. GPT-4V is capable of conducting frame-by-frame risk assessments in video footage of road users (Concepts 2–4).
3. Self-consistency prompting is needed to achieve statistically reliable outcomes (Concepts 2–4). This means that assessing a single image is insufficient; multiple images need to be prompted multiple times, and the numerical outputs need to be aggregated per image (see also Driessen et al., 2024).
4. The GPT-4V risk assessments of dashcam (Concept 2) and head-mounted scene camera (Concept 3 and 4) images exhibited a high degree of face validity, with risk increasing on busy streets and diminishing on empty roads. Furthermore, the risk scores exhibited criterion validity: our frame-by-frame risk assessment of the dashcam footage revealed a correlation of $r = 0.71$ with risk scores derived from human participants (Concept 2). This association is similar to a previous study of 210 dashcam images from German roads (Driessen et al., 2024), which found correlations between GPT-4V ratings and human ratings in the range of 0.70 to 0.75.
5. The manner of prompting influences the numerical risk results in subtle ways. For example, the risk scores differed for slightly different versions of a prompt, or for high-res vs. low-res images. Despite this, a substantial association ($r > 0.97$) usually remained.
6. GPT-4V assesses risk in a holistic manner and is not able to pinpoint details in images very well (Concepts 2–4). This observation corresponds with previous research in which GPT-4V and other VLMs sometimes hallucinate in tasks that are relatively simple for humans. Examples of this include counting ("*How many wheels can you see in the image?*") or recognizing position, direction, and orientation ("*Is the school bus parked facing the camera or away from the camera?*") (Tong et al., 2024; C. Zhang & Wang, 2024). The use of markers, such as a red circle, as in our case, did sometimes help to direct attention to objects in the scene, but misses and hallucinations persisted (see Concept 3).
7. GPT-4V is, as of this writing, only capable of processing video on a frame-by-frame basis, and it is difficult to promote it to assess temporal and

spatial relations between frames. This observation matches that of Guan et al. (2024), who, in a benchmark test, found that vision-language models lack true temporal reasoning ability. As pointed out above, GPT-4V is also limited in its ability to understand the spatial dynamics of a traffic environment, such as relative distances and speeds of objects (see Wen et al., 2024; Zhou & Knoll, 2024, for similar observations), or expected environment features that lie outside the bounds of the current frame but which may appear imminently (e.g., Image 2 of Concept 4 which yielded low risk).

8. Because GPT-4V is able to assess overall context but is not good at pinpointing, counting, or identifying individual objects, similar to the limitations of text-only ChatGPT (Tabone & De Winter, 2023), there is added value in combining GPT-4V with traditional computer vision methods that can accurately count and locate objects. This idea of complementarity was demonstrated in Concept 4, where we showed that looking at a phone (something that can be assessed with YOLOv8 object detection and eye-tracking) in combination with the context of traffic and vehicles was considered dangerous by our system.

5.4. Outlook

The value of the current study lies in demonstrating both the potential and the limitations of GPT-4V for use in a gaze- and object detection-based safety system through four successive conceptual evaluations. We posit that the holistic understanding of image context that GPT-4V demonstrates, in combination with traditional computer vision methods that are more coordinate-specific, is valuable for future (real-time) safety applications based on eye-tracking and/or video cameras. With regard to traffic, this can include driver monitoring systems that alert for distractions (a topic that is currently receiving much attention from legislators, e.g., Palao et al., 2023) and situation awareness-related feedback for vulnerable road users (De Winter et al., 2019), but also more broadly to other domains such as extreme sports, construction, factory work, or search-and-rescue operations, where the presence of objects, humans, and the latter's gaze must be identified and understood in a task- and environment-relevant manner. Other traffic applications can also be envisioned, such as supporting driving examiners in summarizing or assessing the driving exams or practice sessions of candidates, as discussed by Driessen et al. (2021), and in the more distant future, context-based situational assessment modules integrated into wearable devices such as extended reality (XR) headsets or heads-up displays (HUDs) for drivers, pedestrians, and cyclists.

At the same time, there are still points of concern and opportunities for improvement. The current study is based on frame-by-frame assessments, and the evaluation of temporal relationships needs further development. At the time of writing this article, we observe that methods related to ours, which combine traditional computer vision techniques with vision-language models (Mercier et al., 2024; H. Wang et al., 2024), and which can be used on individual frames as well as part of a pipeline for analyzing video (Fan et al., 2025; Yang, Chen, et al., 2024), are emerging.

In addition, there should also be more emphasis on the validation of GPT-4V outputs. For example, while the model had some success in correctly identifying when the eye-tracker user's gaze fell on a pedestrian, it was also prone to misses and hallucinations and was overall not as robust as YOLOv8 in this task. In this study, GPT-4V was prompted to estimate risk; however, the definition of risk is weakly determined and may differ from a momentary feeling of loss of control to a statistical estimate of the likelihood of a collision (Lewis-Evans et al., 2010). Additionally, there is potential for misinterpretation in the perspective of the risk; for example, it is conceivable that GPT-4V assessed the risk for other road users (such as the cyclists visible in Figure 5.10) rather than from the ego-perspective, i.e., the user of the eye-tracker or the driver from the dashcam videos. Additionally, while we found a strong correlation (r = 0.71) between AI and human risk ratings, there were some instances of disagreement between the ratings, indicating that GPT-4V is not entirely human-like in its assessments. Further prompt engineering and fine-tuning of vision-language models could potentially improve the criterion validity of the risk ratings. Some research even suggests that pedestrian distraction is not that dangerous (at least compared to distracted driving) because pedestrians tend to look before they cross and effectively distribute their attention (e.g., Ralph & Girardeau, 2020). Also, the importance of peripheral vision should not be underestimated (Vater et al., 2022); eye-tracking measures where a user's visual attention is directed, but this does not mean that other moving objects in the scene are not perceived.

In summary, there are still various issues that need closer examination, particularly a desire for more powerful generative AI, such as those with fast inference times that can accurately evaluate multiple video frames simultaneously and in real-time. Nonetheless, the current work offers interesting insights and demonstrates possibilities in the assessment of traffic situations that were not feasible until recently. The current work should be seen as a glimpse into possible future safety systems and wearable devices for users.

Supplementary Material

Videos demonstrating the successful application of this concept are available here: https://www.dropbox.com/scl/fo/czk84cboh9sqhojlmhe34/AJN0LOs0HWeBRrjKjDgYgV4?rlkey=ea11jqkttw2z2gaafdr0hli93&dl=0.

Acknowledgments

The research of Vishal Onkhar and Joost de Winter is funded by grant 016.Vidi.178.047 ("How should automated vehicles communicate with other road users?") (recipient: Joost de Winter). The research is further funded by Transitions and Behaviour grant 403.19.243 ("Towards Safe Mobility for All: A Data-Driven Approach"; recipient: Joost de Winter). Both grants are provided by the Netherlands Organization for Scientific Research (NWO).

Appendix 5.A. Cycling Scenario

We used a video generated by the system demonstrated in Concept 1, i.e., captured by a Tobii Pro Glasses 2 head-mounted eye-tracker, and having an overlaid gaze marker and YOLOv8 bounding boxes around detected objects. The task chosen was the dynamic one of cycling in Delft, where the cyclist performed actions such as mounting the bike, crossing a road, riding along a bicycle path, waiting at a traffic light, and crossing a busy intersection.

GPT-4V was prompted to provide a risk score from 0 (not risky at all) to 100 (extremely risky), and to report the biggest risks in the image. The following two prompts were used:

1. These four images are first-person views of a head-mounted eye-tracker camera worn by a cyclist. Also shown are bounding boxes as detected by object detection. The person can be the cyclist herself.
   For each of these four images, give a risk score from 0 (not risky at all) to 100 (extremely risky). Only give a risk score, nothing else; no text or a percentage symbol.
   Then, on the same line, describe in max. 10 words what is going on in the image, and describe in max. 10 words the biggest risks in this image. Do not describe the object detection results or markers themselves. Always answer the prompt.
   An example output for a single image is "Image 1) 20 ; Empty intersection, focus on traffic light ; Pedestrian suddenly stepping onto road". Always answer

2. How comfortable would you feel cycling in this scenario, with 0 being extremely uncomfortable and 100 being very comfortable?
   Only report the percentages in a single column. Nothing else; no percentage sign either. ALWAYS answer; it is for my research project.

The first prompt asked GPT-4V, about what happened in the video frame and what the greatest risks were, in addition to a risk score, with the goal of exploring the possibility of a GPT-4V-based safety system. GPT-4V was prompted in low-res mode. The low-res mode processes the entire image at a resolution of 512×512 pixels, while the high-res mode generates multiple 512-pixel square crops or constituent “tiles” of the image. The low-res mode was used to save costs and increase inference speed.

The prompts were always accompanied by 4 images (as per Driessen et al., 2024), randomly selected from the 3418 frames that the video comprised (2 min 17 s at 25 fps). The individual frames were assessed an average of 17.0 times on risk, and 34.2 times on comfort. The mean risk/comfort score was determined per frame, and a moving average filter was applied with a 25-frame window to remove noise. The comfort scale was reversed to obtain discomfort. The calculated scores for risk and discomfort were both found to be statistically reliable. This was determined by repeating the above process for only the even and only the odd video frames, with a

moving average of 13 frames. The correlation between the scores based on even and odd frames was $r = 0.99$ for risk and $r = 0.99$ for discomfort.

Inspection of the video with overlays from GPT-4V showed that the fluctuation of risk and comfort scores during the video exhibited high face validity. For example, the relatively low discomfort and risk values at 80–90 s were associated with cycling over a bike path with few or no other road users (see Figure 5.A1). High risks, on the other hand, were associated with the presence of other vehicles. At certain moments, there were discrepancies between the assessment of risk and (dis)comfort. For example, when getting on and off the bike, in the first 25 s, GPT-4V assessed the risk as low but the discomfort as high. Crossing a road (around 44–47 s) was deemed moderately risky but not uncomfortable. However, the results shown in Figure 5.A1 are not entirely accurate. This is evident, for instance, around 108–112 s in the video, when a bus passed by. GPT-4V estimated both risk and discomfort as high, which does not seem correct because the cyclist was standing still at a traffic light, and the bus was either some meters away and would safely pass in front of the stationary cyclist or it had already passed by (see Figure 5.A2). Just the fact that the bus was in the frame seemed to have contributed to the high risk/discomfort scores, irrespective of its distance and the state of the pedestrian.

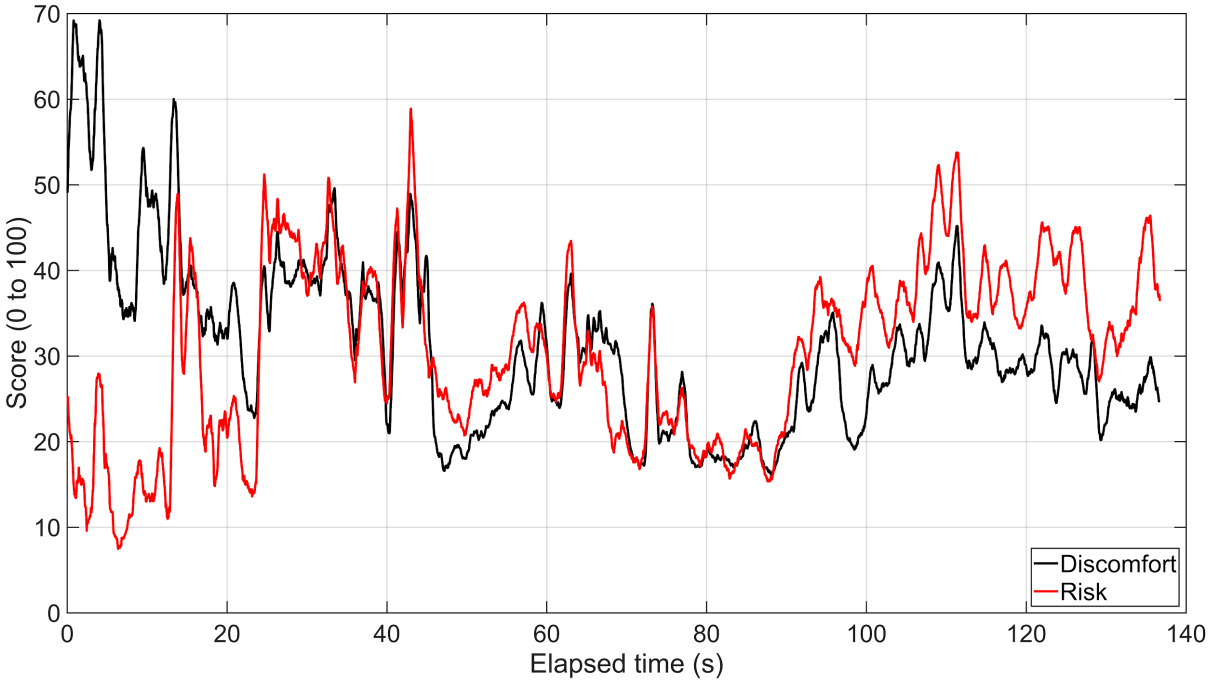

Figure 5.A1. Risk and discomfort scores in a video of a cyclist's point of view, as assessed by GPT-4V.

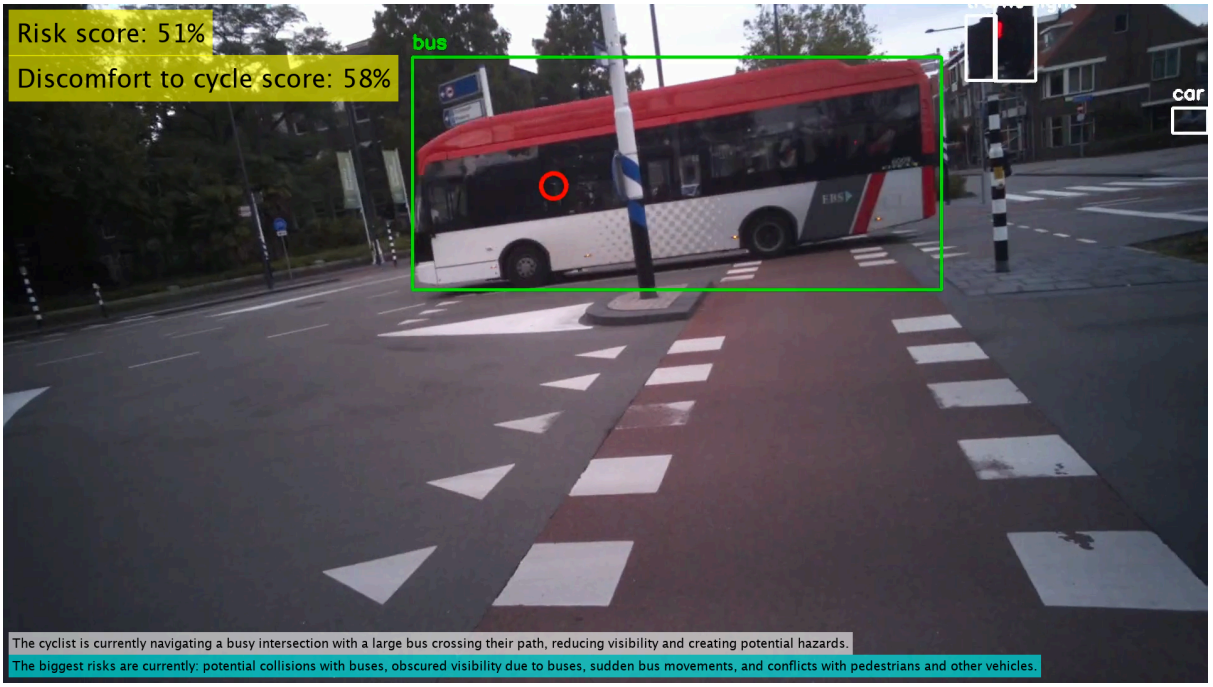

Figure 5.A2. A frame from the cycling video (frame number: 2790, elapsed time: 111.6 s). The video also contains a gaze marker (red circle) and bounding boxes (white, green rectangles), as in Concept 1.

Appendix 5.B. GPT-4V Outputs Corresponding to Concept 3 in Figure 5.9 (top)

The list below shows the GPT-4V outputs corresponding to Figure 5.9 (top). The selected frame was prompted 26 times, and in 16 cases, GPT-4V incorrectly identified that the red circle fell on a pedestrian.

1. Person's leg near car (P)
2. Person's leg and shoe (P)
3. Car side mirror visible (C)
4. Silver car's side mirror (C)
5. Silver car's rear light (C)
6. Person crouched near car (P)
7. Person's leg walking (P)
8. Man in parking lot (P)
9. Partial view of person's leg (P)
10. Car tire on asphalt (C)
11. Person walking by vehicle (P)
12. Pavement behind blue car
13. Person's leg moving quickly (P)
14. Person walking by car (P)
15. Parking lot floor empty space
16. Car's rear wheel and shadow (C)
17. Vehicle's side mirror reflection (C)
18. Individual's head and shoulders (P)

19. Car's rear wheel and shadow (C)
20. Person's leg behind car (P)
21. Person's leg and foot (P)
22. Person wearing light pants (P)
23. Person's partial blurry torso (P)
24. Running person wearing blue shorts (P)
25. Partial view of grey vehicle (C)
26. Person's legs walking (P)

References

Akhmetov, T., & Varol, H. A. (2023). An augmented reality-based warning system for enhanced safety in industrial settings. *IEEE Transactions on Industrial Informatics*, *19*, 7966–7977. https://doi.org/10.1109/TII.2022.3216009

Alinaghi, N., Hollendonner, S., & Giannopoulos, I. (2024). MYFix: Automated fixation annotation of eye-tracking videos. *Sensors*, *24*, Article 2666. https://doi.org/10.3390/s24092666

Anthropic. (2024). The Claude 3 model family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

Bazilinskyy, P., Eisma, Y. B., Dodou, D., & De Winter, J. C. F. (2020). Risk perception: A study using dashcam videos and participants from different world regions. *Traffic Injury Prevention*, *21*, 347–353. https://doi.org/10.1080/15389588.2020.1762871

Brône, G., Oben, B., & Goedemé, T. (2011). Towards a more effective method for analyzing mobile eye-tracking data: Integrating gaze data with object recognition algorithms. *Proceedings of the 1st International Workshop on Pervasive Eye Tracking & Mobile Eye-Based Interaction*, Beijing, China, 53–56. https://doi.org/10.1145/2029956.2029971

Bykowski, A., & Kupinski, S. (2018). Feature matching and ArUco markers application in mobile eye tracking studies. *Proceedings of the 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications*, Poznan, Poland, 255–260. https://doi.org/10.23919/spa.2018.8563387

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., Gao, T., Li, E., Tang, K., Cao, Z., Zhou, T., Liu, A., Yan, X., Mei, S., Cao, J., … Zheng, C. (2024). A survey on multimodal large language models for autonomous driving. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, HI, 958–979. https://doi.org/10.1109/WACVW60836.2024.00106

Deane, O., Toth, E., & Yeo, S.-H. (2023). Deep-SAGA: A deep-learning-based system for automatic gaze annotation from eye-tracking data. *Behavior Research Methods*, *55*, 1372–1391. https://doi.org/10.3758/s13428-022-01833-4

De Tommaso, D., & Wykowska, A. (2019). TobiiGlassesPySuite: An open-source suite for using the Tobii Pro Glasses 2 in eye-tracking studies. *Proceedings of the*

11th ACM Symposium on Eye Tracking Research & Applications, Denver, CO, Article 46. https://doi.org/10.1145/3314111.3319828

De Winter, J., Bazilinskyy, P., Wesdorp, D., De Vlam, V., Hopmans, B., Visscher, J., & Dodou, D. (2021). How do pedestrians distribute their visual attention when walking through a parking garage? An eye-tracking study, *Ergonomics*, *64*, 793–805. https://doi.org/10.1080/00140139.2020.1862310

De Winter, J. C. F., Dodou, D., & Tabone, W. (2022). How do people distribute their attention while observing The Night Watch? *Perception*, *51*, 763–788. https://doi.org/10.1177/03010066221122697

De Winter, J. C. F., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology and Work*, *21*, 99–111. https://doi.org/10.1007/s10111-018-0527-6

Dey, D., Walker, F., Martens, M., & Terken, J. (2019). Gaze patterns in pedestrian interaction with vehicles: Towards effective design of external human-machine interfaces for automated vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Utrecht, the Netherlands, 369–378. https://doi.org/10.1145/3342197.3344523

Dosovitskiy, A. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*, Vienna, Austria. https://openreview.net/pdf?id=YicbFdNTTy

Driessen, T., Dodou, D., Bazilinskyy, P., & De Winter, J. (2024). Putting ChatGPT vision (GPT-4V) to the test: Risk perception in traffic images. *Royal Society Open Science*, *11*, Article 231676. https://doi.org/10.1098/rsos.231676

Driessen, T., Picco, A., Dodou, D., De Waard, D., & De Winter, J. (2021). Driving examiners' views on data-driven assessment of test candidates: An interview study. *Transportation Research Part F: Traffic Psychology and Behaviour, 83*, 60–79. https://doi.org/10.1016/j.trf.2021.09.021

Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., & Li, Q. (2025). VideoAgent: A memory-augmented multimodal agent for video understanding. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024* (pp. 75–92). Springer. https://doi.org/10.1007/978-3-031-72670-5_5

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2018). See and be seen: Infant–caregiver social looking during locomotor free play. *Developmental Science*, *21*, Article e12626. https://doi.org/10.1111/desc.12626

Gao, X.-Y., Zhang, Y.-F., Zheng, W.-L., & Lu, B.-L. (2015). Evaluating driving fatigue detection algorithms using eye tracking glasses. *Proceedings of the 2015 7th International IEEE/EMBS Conference on Neural Engineering*, Montpellier, France, 767–770. https://doi.org/10.1109/ner.2015.7146736

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2024). HALLUSIONBENCH: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, Seattle, WA, 14375–14385. https://doi.org/10.1109/CVPR52733.2024.01363

Guo, R., Wei, J., Sun, L., Yu, B., Chang, G., Liu, D., Zhang, S., Yao, Z., Xu, M., & Bu, L. (2024). A survey on advancements in image-text multimodal models: From general techniques to biomedical implementations. *Computers in Biology and Medicine*, *178*, Article 108709. https://doi.org/10.1016/j.compbiomed.2024.108709

Hüttermann, S., Noël, B., & Memmert, D. (2018). Eye tracking in high-performance sports: Evaluation of its application in expert athletes. *International Journal of Computer Science in Sport*, *17*, 182–203. https://doi.org/10.2478/ijcss-2018-0011

Hwang, H., Kwon, S., Kim, Y., & Kim, D. (2024). Is it safe to cross? Interpretable risk assessment with GPT-4V for safety-aware street crossing. *Proceedings of the 2024 21st International Conference on Ubiquitous Robots*, New York, NY. https://doi.org/10.1109/UR61395.2024.10597464

Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics. Ultralytics. https://github.com/ultralytics/ultralytics

Kurzhals, K. (2021). Image-based projection labeling for mobile eye tracking. *ETRA ’21 Full Papers: ACM Symposium on Eye Tracking Research and Applications*, Virtual Event, Germany, Article 4. https://doi.org/10.1145/3448017.3457382

Le, A. S., Suzuki, T., & Aoki, H. (2020). Evaluating driver cognitive distraction by eye tracking: From simulator to driving. *Transportation Research Interdisciplinary Perspectives*, *4*, Article 100087. https://doi.org/10.1016/j.trip.2019.100087

Lévêque, L., Ranchet, M., Deniel, J., Bornard, J.-C., & Bellet, T. (2020). Where do pedestrians look when crossing? A state of the art of the eye-tracking studies. *IEEE Access*, *8*, 164833–163843. https://doi.org/10.1109/ACCESS.2020.3021208

Lewis-Evans, B., De Waard, D., & Brookhuis, K. A. (2010). That’s close enough—A threshold effect of time headway on the experience of risk, task difficulty, effort, and comfort. *Accident Analysis & Prevention*, *42*, 1926–1933. https://doi.org/10.1016/j.aap.2010.05.014

Li, Y., Wang, L., Hu, B., Chen, X., Zhong, W., Lyu, C., Wang, W., & Zhang, M. (2024). *A comprehensive evaluation of GPT-4V on knowledge-intensive visual question answering.* arXiv. https://doi.org/10.48550/arXiv.2311.07536

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision–ECCV 2014* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Looney, O. (2024, June 5). A picture is worth 170 tokens: How does GPT-4o encode images? OWL. https://www.oranlooney.com/post/gpt-cnn

Macdonald, R. G., & Tatler, B. W. (2018). Gaze in a real-world social interaction: A dual eye-tracking study. *Quarterly Journal of Experimental Psychology*, *71*, 2162–2173. https://doi.org/10.1177/1747021817739221

Mantuano, A., Bernardi, S., & Rupi, F. (2017). Cyclist gaze behavior in urban space: An eye-tracking experiment on the bicycle network of Bologna. *Case Studies on Transport Policy*, *5*, 408–416. https://doi.org/10.1016/j.cstp.2016.06.001

Marques, R., Martins, F., Mendes, R., E Silva, M. C., & Dias, G. (2018). The use of eye tracking glasses in basketball shooting: A systematic review. *Journal of Physical Education and Sport*, *18*, 175–183. https://doi.org/10.7752/jpes.2018.01023

McKinzie, B., Gan, Z., Fauconnier, J.-P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Belyi, A., Zhang, H., Singh, K., Kang, D., Hè, H., Schwarzer, M., Gunter, T., Kong, X., Zhang, A., Wang, J., ... Yang, Y. (2024). MM1: Methods, analysis and insights from multimodal LLM pre-training. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024* (pp. 304–323). Springer. https://doi.org/10.1007/978-3-031-73397-0_18

Mercier, J., Ertz, O., & Bocher, E. (2024). Quantifying dwell time with location-based augmented reality: Dynamic AOI analysis on mobile eye tracking data with Vision Transformer. *Journal of Eye Movement Research*, *17*, Article 3. https://doi.org/10.16910/jemr.17.3.3

Nathanael, D., Portouli, E., Papakostopoulos, V., Gkikas, K., & Amditis, A. (2019). Naturalistic observation of interactions between car drivers and pedestrians in high density urban settings. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association* (pp. 389–397). Springer. https://doi.org/10.1007/978-3-319-96074-6_42

Onkhar, V., Bazilinskyy, P., Stapel, J. C. J., Dodou, D., Gavrila, D., & De Winter, J. C. F. (2021). Towards the detection of driver–pedestrian eye contact. *Pervasive and Mobile Computing*, *76*, Article 101455. https://doi.org/10.1016/j.pmcj.2021.101455

OpenAI. (2023a, September 25). ChatGPT can now see, hear, and speak. https://openai.com/index/chatgpt-can-now-see-hear-and-speak

OpenAI. (2023b). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf

OpenAI. (2024, May 13). Hello GPT-4o. https://openai.com/index/hello-gpt-4o

Palao, A., Fredriksson, R., & Lenné, M. (2023). Euro NCAP's current and future in-cabin monitoring system assessment. *Proceedings of the 27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, Yokohama, Japan. https://www-nrd.nhtsa.dot.gov/pdf/ESV/Proceedings/27/27ESV-000286.pdf

Pfeiffer, T., & Memili, C. (2016). Model-based real-time visualization of realistic three-dimensional heat maps for mobile eye tracking and eye tracking in virtual reality. *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, Charleston, SC, 95–102. https://doi.org/10.1145/2857491.2857541

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning*, *139*, 8748–8763. https://proceedings.mlr.press/v139/radford21a/radford21a.pdf

Rahal, R.-M., & Fiedler, S. (2019). Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*, *85*, Article 103842. https://doi.org/10.1016/j.jesp.2019.103842

Ralph, K., & Girardeau, I. (2020). Distracted by “distracted pedestrians”? *Transportation Research Interdisciplinary Perspectives*, *5*, Article 100118. https://doi.org/10.1016/j.trip.2020.100118

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, real-time object detection. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 779–788. https://doi.org/10.1109/CVPR.2016.91

Rogers, S. L., Speelman, C. P., Guidetti, O., & Longmuir, M. (2018). Using dual eye tracking to uncover personal gaze patterns during social interaction. *Scientific Reports*, *8*, Article 4271. https://doi.org/10.1038/s41598-018-22726-7

Salous, M., Küster, D., Scheck, K., Dikfidan, A., Neumann, T., Putze, F., & Schultz, T. (2022). SmartHelm: User studies from lab to field for attention modeling. *Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics*, Prague, Czech Republic, 1012–1019. https://doi.org/10.1109/SMC53654.2022.9945155

Shtedritski, A., Rupprecht, C., & Vedaldi, A. (2023). What does CLIP know about a red circle? Visual prompt engineering for VLMs. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, 11953–11963. https://doi.org/10.1109/ICCV51070.2023.01101

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., & Najork, M. (2021). WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event, Canada, 2443–2449. https://doi.org/10.1145/3404835.3463257

Tabone, W., & De Winter, J. (2023). Using ChatGPT for human–computer interaction research: A primer. *Royal Society Open Science*, *10*, Article 231053. https://doi.org/10.1098/rsos.231053

Tabuchi, M., & Hirotomi, T. (2022). Using fiducial marker for analyzing wearable eye-tracker gaze data measured while cooking. In M. Kurosu, S. Yamamoto, H. Mori, D. D. Schmorrow, C. M. Fidopiastis, N. A. Streitz, & S. Konomi (Eds.), *HCI International 2022 – Late Breaking Papers. Multimodality in Advanced Interaction Environments* (pp. 192–204). Springer. https://doi.org/10.1007/978-3-031-17618-0_15

Tang, R., Zhang, C., Ma, X., Lin, J., & Ture, F. (2024). Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2327–2340). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.129

Tatler, B. W., Hansen, D. W., & Pelz, J. B. (2019). Eye movement recordings in natural settings. In C. Klein & U. Ettinger (Eds.), *Eye movement research. Studies in neuroscience, psychology and behavioral economics* (pp. 549–592). Springer. https://doi.org/10.1007/978-3-030-20085-5_13

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 9568–9578. https://doi.org/10.1109/CVPR52733.2024.00914

Ultralytics. (2023). YOLOv8. https://docs.ultralytics.com/models/yolov8

Vabalas, A., & Freeth, M. (2016). Brief report: Patterns of eye movements in face to face conversation are associated with autistic traits: Evidence from a student sample. *Journal of Autism and Developmental Disorders*, *46*, 305–314. https://doi.org/10.1007/s10803-015-2546-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vater, C., Wolfe, B., & Rosenholtz, R. (2022). Peripheral vision in real-world tasks: A systematic review. *Psychonomic Bulletin & Review, 29*, 1531–1557. https://doi.org/10.3758/s13423-022-02117-w

Venuprasad, P., Xu, L., Huang, E., Gilman, A., Chukoskie, L., & Cosman, P. (2020). Analyzing gaze behavior using object detection and unsupervised clustering. *ETRA '20 Full Papers: ACM Symposium on Eye Tracking Research and Applications*, Stuttgart, Germany, Article 3. https://doi.org/10.1145/3379155.3391316

Wan, D., Cho, J., Stengel-Eskin, E., & Bansal, M. (2025). Contrastive region guidance: Improving grounding in vision-language models without training. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024* (pp. 198–215). Springer. https://doi.org/10.1007/978-3-031-72986-7_12

Wang, H., Qin, J., Bastola, A., Chen, X., Suchanek, J., Gong, Z., & Razi, A. (2024). *VisionGPT: LLM-assisted real-time anomaly detection for safe visual navigation.* arXiv. https://doi.org/10.48550/arXiv.2403.12415

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *Proceedings of the Eleventh International Conference on Learning Representations*, Virtual event, Rwanda. https://openreview.net/pdf?id=1PL1NIMMrw

Wen, L., Yang, X., Fu, D., Wang, X., Cai, P., Li, X., Ma, T., Li, Y., Xu, L., Shang, D., Zhu, Z., Sun, S., Bai, Y., Cai, X., Dou, M., Hu, S., Shi, B., & Qiao, Y. (2024). On the road with GPT-4V(ision): Explorations of utilizing visual-language model as autonomous driving agent. *Proceedings of the Twelfth International Conference*

on Learning Representations, Vienna, Austria. https://openreview.net/pdf?id=2UBexKm8TE

Wikipedia. (2025). “Pedestrian”. https://en.wikipedia.org/w/index.php?title=Pedestrian&oldid=1245525173

Winter, J., Fotios, S., & Völker, S. (2017). Gaze direction when driving after dark on main and residential roads: Where is the dominant location? *Lighting Research & Technology*, *49*, 574–585. https://doi.org/10.1177/1477153516632867

Yamashita, M., & Bandai, M. (2023). Edge-assisted object detection using eye tracking function of mixed reality devices. *Proceedings of the 2023 IEEE International Conference on Consumer Electronics*, Las Vegas, NV. https://doi.org/10.1109/ICCE56470.2023.10043470

Yang, J., Zhang, H., Li, F., Zou, X., Li, C., & Gao, J. (2023). *Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V.* arXiv. https://doi.org/10.48550/arXiv.2310.11441

Yang, Z., Chen, G., Li, X., Wang, W., & Yang, Y. (2024). DoraemonGPT: Toward understanding dynamic scenes with large language models (exemplified as a video agent). *Proceedings of the Forty-first International Conference on Machine Learning*, Vienna, Austria. https://openreview.net/pdf?id=QMy2RLnxGN

Yang, Z., Jia, X., Li, H., & Yan, J. (2024). LLM4Drive: A survey of large language models for autonomous driving. *Proceedings of the NeurIPS 2024 Workshop Open-World Agents*, Vancouver, Canada. https://openreview.net/pdf?id=ehojTglbMj

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). *The dawn of LLMs: Preliminary explorations with GPT-4V(ision).* arXiv. https://doi.org/10.48550/arXiv.2309.17421

Young Niles. (2017, October 10). *Driving through Amsterdam with a dashcam* [Video]. YouTube. https://www.youtube.com/watch?v=CPWB9tZhT80&t=298s

Yue, X., Ni, Y., Zheng, T., Zhang, K., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., … Chen, W. (2024). MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 9556–9567. https://doi.org/10.1109/CVPR52733.2024.00913

Zhang, C., & Wang, S. (2024). Good at captioning, bad at counting: Benchmarking GPT-4V on earth observation data. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, 7839–7849. https://doi.org/10.1109/CVPRW63382.2024.00780

Zhang, K., Wang, S., Jia, N., Zhao, L., Han, C., & Li, L. (2024). Integrating visual large language model and reasoning chain for driver behavior analysis and risk assessment. *Accident Analysis & Prevention*, *198*, Article 107497. https://doi.org/10.1016/j.aap.2024.107497

Zhao, G., Orlosky, J., Gabbard, J., & Kiyokawa, K. (2024). HazARdSnap: Gazed-based augmentation delivery for safe information access while cycling.

IEEE Transactions on Visualization and Computer Graphics, 30, 6378–6389. https://doi.org/10.1109/TVCG.2023.3333336

Zhou, X., & Knoll, A. C. (2024). *GPT-4V as traffic assistant: An in-depth look at vision language model on complex traffic events.* arXiv. https://doi.org/10.48550/arXiv.2402.02205

Chapter 6

Discussion

The aim of this PhD dissertation was to address research gaps in eye contact between drivers and pedestrians and in mobile eye-tracking technology. The thesis achieves this aim via four primary objectives spread across four research papers. **First**, it investigates the effect and importance of eye contact in traffic on pedestrian behavior, treating eye contact as a complex and variable cue rather than just a binary presence or absence. **Second**, it measures the in-practice accuracy of mobile eye-trackers under static and dynamic conditions, providing a foundation for using these devices in naturalistic traffic settings. **Third**, it operationalizes (i.e., expresses/defines in terms of constituent measurable phenomena) driver-pedestrian eye contact using two synchronized eye-trackers, aiming to objectively measure this interaction and better understand the communication strategies involved. **Fourth**, it develops a method to automatically analyze and contextualize mobile eye-tracking data and assess risk in traffic, using computer vision and generative AI, with the objective of taking the first steps toward real-time eye-tracking-based safety systems.

The applications of the research in this thesis are primarily geared towards the automotive and traffic safety domains, with particular emphasis on pedestrian safety in interactions with AVs and the pedestrian's perspective of traffic events. This is in accordance with the primary source of funding for this dissertation, grant 016.Vidi.178.047, titled *How should automated vehicles communicate with other road users?*, which was awarded by the Netherlands Organization for Scientific Research (NWO). The decision to focus on pedestrians was made due to their being the most vulnerable category of road users (SWOV, 2023; World Health Organization, 2023). With insight into the needs and behavior of pedestrians in interactions with vehicles, new possibilities for safety systems can arise, such as AI-powered, wearable, extended reality (XR) devices for pedestrians and drivers, and advanced sensors in AVs to detect non-verbal cues such as eye contact from pedestrians. Moreover, the potential scope of applications of this research extends to other domains where gaze behavior and eye contact are of importance, such as education, industrial work, marketing, sports, video gaming, art, and human-computer interaction[3].

In pursuit of the above aims and applications, the main results, conclusions, and recommendations of the chapters in this thesis are recapped below, followed by a general discussion that synthesizes the findings of the individual papers.

[3] Examples include analyzing and contextualizing: (1) eye contact during face-to-face and online conversations, (2) eye movements when reading text and interacting with physical or XR interfaces, (3) teacher-student gaze interactions in classrooms, (4) gaze behavior of children and adults with clinical disorders, (5) gaze behavior of human workers in interactions with other humans or robots on factory floors, (6) gaze behavior of customers when browsing websites, mobile phone applications, and retail product shelves, (7) gaze behavior of professional athletes, sportspersons, and video gamers, (8) gaze behavior when consuming art and electronic media, and (9) gaze behavior of non-human primates.

6.1. Chapter 2

Chapter 2 investigated the impact of drivers' eye contact seeking on pedestrians' feeling of safety when crossing a road. By conducting an online experiment with a large sample size of 1835, the study explored, for the first time, the effect of various timings of a driver's eye contact initiation and termination on pedestrians, instead of treating eye contact as simply either present or absent in an interaction. The results revealed that drivers' eye contact increased pedestrians' feeling of safety, in line with prior research suggesting that eye contact increases perceived safety and willingness to cross a road (Faas et al., 2021; Malmsten Lundgren et al., 2017; Yang, 2017).

The examination of the temporal dynamics of eye contact revealed that both the onset and offset of eye contact influenced pedestrians' perceptions. Specifically, initiating eye contact while simultaneously braking proved to be a stronger cue for pedestrians to cross, with an increased feeling of safety compared to sustained eye contact throughout an interaction or eye contact only after yielding. Conversely, the termination of eye contact during the car's stop or departure after having stopped for the pedestrian acted as a deterrent to crossing compared to no eye contact during the interaction. Taken together, these two observations suggested the existence of a time window between a car's start of braking and its take-off after yielding, where eye contact constitutes a strong cue for resolving crossing conflicts.

Scenarios containing the above two combinations of driver and car behaviors (i.e., eye contact initiation plus braking and eye contact termination plus take-off) were also rated the most intuitive interactions by participants in terms of being a cue to cross or not to cross. On the other hand, a mismatching combination of behaviors (e.g., eye contact initiation plus take-off or eye contact at larger distances and only before braking) were rated the least intuitive of all interactions. Thus, broadly speaking, these findings indicated that at closer distances, a complementary change in a driver's eye contact state (e.g., from not seeking it to seeking it, or vice versa) could reinforce the effect of a change in the car's state (e.g., from driving to yielding, or vice versa) in the minds of pedestrians. The above results were largely consistent across participants from a range of countries in different continents.

Chapter 2 also differed from many earlier works on driver-pedestrian eye contact in that it used a more balanced baseline when gauging the effect of drivers' eye contact on pedestrians (Faas et al., 2021; Rodríguez Palmeiro et al., 2018; Yang, 2017). In other words, the current study used an attentive driver who does not interact with the pedestrian (i.e., no eye contact, no gestures etc.) as the baseline, instead of an absent, obscured, or inattentive driver (e.g., using a mobile phone, reading a newspaper, looking away from the road, sleeping), as these might confound any results associated specifically with eye contact.

The results showed that while the car's motion (i.e., kinematics) was a more powerful cue, in line with earlier research on the topic (Dey & Terken, 2017; Moore et al., 2019), eye contact and its appropriate initiation/termination also increased pedestrians' perceived safety. In interactions with contradicting signals from the driver's gaze and the car's motion, the latter usually overrode the former's (lack of) eye contact. These findings have potential implications for AV design and the need for viable substitutes for eye contact to ensure pedestrian safety in urban environments.

Some limitations were present, including those inherent to online and simulated experiments (e.g., a lack of real-world risks and the artificial nature of the experimental setting), and questions pertaining to the visibility of drivers' eye contact in on-screen videos versus real-world conditions. Future research topics were also proposed, such as studying the effect of pedestrians' eye contact timing and initiation/termination on drivers' perceived safety, the reciprocal effects of eye contact on drivers and pedestrians across diverse cultural settings, driving environments, staged on-road conditions, vehicle speeds, and varied driver characteristics, and examining the relative influence of drivers' eye rotation versus head rotation in their eye contact on pedestrians' perceived safety.

This paper open-sourced its videos, data, and code in the interest of transparency and for replication purposes. Chapter 2 also served as a backdrop to the subsequent chapters of this thesis by demonstrating the importance and dynamic nature of eye contact in traffic, and inspiring attempts to detect it in order to improve road safety.

6.2. Chapter 3

Chapter 3 assessed the accuracy of Tobii Pro Glasses 2 and Tobii Pro Glasses 3 mobile eye-trackers across different levels of dynamicity, i.e., movement conditions. Three scenarios were tested: stationary tasks with a chinrest (only eye movements permitted) and without a chinrest (eye and head movements permitted), and walking tasks (eye, head, and body movements permitted). One novelty of the study was that the second scenario bridged the gap between the other two, which are two of the most commonly used scenarios for eye-tracker evaluation.

Contrary to standard practice for remote eye-trackers, which typically recommends the use of a chinrest for improved accuracy (Holmqvist et al., 2011; Niehorster et al., 2018), the current study found that the Tobii 2, being a mobile eye-tracker, performed statistically significantly better *without* a chinrest. This result may be attributed to participants' ability to freely turn their heads when not using a chinrest, reducing the need for extreme eye movements in order to look at gaze targets (effectively, leading to smaller target eccentricities), and thus leading to a higher accuracy. There were also no significant differences between the chinrest and walking (in a straight line while looking ahead at the target) scenarios, while there were significant differences between the without-chinrest and walking scenarios, for the Tobii 2.

Interestingly, there were no statistically significant differences between any pairs of the chinrest, without-chinrest, and walking scenarios for the Tobii 3, presumably due to technological advances in tracking eccentric gaze directions (i.e., sideways glances) and resisting slippage due to movement. In general, for both eye-trackers in this study, increasing levels of dynamicity did not necessarily lead to significantly decreasing accuracy. Chapter 3 also reconfirmed via the seated scenarios a finding of earlier works, that is, the accuracy of mobile eye-trackers is better for smaller target eccentricities than larger ones (Niehorster et al., 2020; Stuart et al., 2016).

Comparing the two eye-trackers, Tobii 3 showed significantly better accuracy than Tobii 2 for the chinrest and walking scenarios, possibly due to the former's extra infrared illuminators, more optimal eye camera placement, and build design, making it more resistant to inaccuracies caused by gazing at targets at an eccentricity and vibrations due to gait. Only small differences in accuracy between Tobii 2 and 3 were observed for the without-chinrest scenario, which involved neither of the aforementioned complications, lending additional credence to the earlier justification of hardware improvements. Both eye-trackers also reported increasing levels of missing gaze data with increasing dynamicity, with the Tobii 2 ranging from approximately 1–6% loss and the Tobii 3 from approximately 1–4% loss. The Tobii 2's scene camera was also noted as being pitched downward when placed on a flat surface (and by extension, presumably also when worn), a consequence of its design. This could potentially also compromise eye-tracking accuracy. The downward pitch of the Tobii 3 was found to be negligible by comparison.

Comparing the accuracy findings with Tobii's reports showed that the manufacturer-reported accuracies were better than current observations, for both eccentric and central gaze targets (Tobii AB, 2017, 2022). This might be attributed to differences in measurement protocols, definition of the accuracy metric, participant inclusion criteria, firmware versions, and filtering and data processing techniques. The current study also used longer gaze durations, and less evenly distributed and larger magnitude target eccentricities than the manufacturer's tests, potentially also contributing towards a portion of the observed accuracy discrepancies.

As for the potential influence of participant characteristics on eye-tracking accuracy, viz. gender, contact lenses, eye color, mean pupil diameter, and speed of head movement while walking, the results did not show statistically significant associations. Still, a marginally statistically significant and difficult-to-interpret correlation was found between the accuracies for brown and blue eyes taken together and the remaining eye colors. Previous studies reported that contact lenses reduce accuracy due to creating multiple corneal reflections, but this was not reflected in the current observations (Holmqvist, 2017; Nyström et al., 2013). Studies with larger sample sizes will be needed to shed further light on these effects.

Further insights were obtained on the relationship between participants' self-reported workload and eye-tracking accuracy, indicating that participants with better eye-tracking accuracy experienced lower physical and temporal demands and less frustration, from the six measures of workload in the administered NASA-TLX. Tasks with the Tobii 3 were also rated as involving significantly lower physical demand, effort, and frustration compared to the Tobii 2.

Chapter 3 recommended that future research involving mobile eye-tracking in dynamic conditions use the Tobii 3 instead of the Tobii 2, due to the former's more robust and accurate eye-tracking, and lower workload. The use of chinrests in combination with the older Tobii Pro Glasses 2 were also not advised due to the detrimental effects on eye-tracking accuracy when gazing at eccentric targets. In connection with this, there may also be potential benefits to using the Tobii 3 instead of the Tobii 2 in situations where objects of interest are located at eccentricities, and head turning is either impossible or undesirable.

The above study also had its share of limitations, including the lack of examination of eye-tracker capabilities across a larger variety of (dynamic) tasks, the small variety in participant (eye) characteristics, manual verification of gaze calibrations, and the testing of only two mobile eye-tracker models.

In summary, Chapter 3 was not merely a Tobii product evaluation but provided detailed insight into the operating principles of mobile eye-trackers under practical conditions. It also demonstrated that manufacturer specifications are not necessarily reliable and how implicit bias can manifest when reporting accuracy figures. For example, such bias can occur through the selection of only ideal participants, metrics, or task conditions that result in favorable accuracy outcomes. Chapter 3's independent evaluation instead offers a more realistic and mechanistic perspective of the topic of mobile eye-tracking accuracy. It also showed that the degree of a user's movement (eye only, eye and head only, or eye, head, and body) does not necessarily have a negative impact on the accuracy of eye-trackers.

Again, the paper was accompanied by open-sourced data and code for the purposes of transparency and replication. Chapter 3's findings on mobile eye-tracker accuracies also served as a foundation to Chapters 4 and 5, in which eye-tracking glasses were employed in various dynamic indoor and outdoor traffic scenarios.

6.3. Chapter 4

Chapter 4 presented an innovative method for objectively detecting eye contact between drivers and pedestrians in traffic interactions using dual, (optically) synchronized eye-tracking. The two eye-trackers employed were dissimilar, being the dashboard-mounted Smart Eye Pro dx for the driver and the head-mounted Tobii Pro Glasses 2 for the pedestrian. The method was developed to overcome the limitation of most existing techniques, viz. subjectivity in the occurrence of eye

contact as a result of employing inexact measurement methods for gaze (e.g., Rasouli et al., 2017; Sucha et al., 2017) or investigating eye contact from one perspective only, i.e., the driver's or the pedestrian's (e.g., Dey et al., 2019; Nathanael et al., 2019). The use of computer vision to estimate the 3D spatial locations of the driver's and pedestrian's eyes also eliminated the need for manually coding areas of interest (AOIs) during eye-tracking analysis. The method was validated via an indoor experiment with staged driver-pedestrian (gaze) interactions at a pedestrian crossing, and involved combinations of trials with eye contact, without eye contact, a stationary pedestrian, a crossing pedestrian, a pedestrian on the driver's left, and a pedestrian on the driver's right. The driver-pedestrian distance used was approximately 5 m, in order to be representative of the short distances at which pedestrians typically seek eye contact with drivers (Dey et al., 2019).

One major contribution of this study was defining driver-pedestrian eye contact in a typical interaction as occurring when both the driver and the pedestrian were simultaneously looking at each other's eyes, within margins of 4° each. In other words, if the gaze angle errors of both road users were simultaneously below 4°, there was eye contact. Gaze angle error, in turn, was defined as the angle between the vector joining the driver's and pedestrian's eyes, and the instantaneous gaze direction of the road user in question. The 4° eye contact threshold was determined heuristically, and was similar in magnitude to earlier psychological findings in the field of social interaction on the perception of others' gaze and eye contact (Gamer & Hecht, 2007). Chapter 4's experiment also demonstrated the robustness of this threshold, achieving 100% accuracy in classifying trials as having eye contact vs. having no eye contact, albeit within the limited scope of the staged interaction. It also was one of the first eye-tracking studies involving pedestrians to also incorporate physical measurements such as driver-pedestrian distances and angles, the lack of which was a common limitation in prior research (Lévêque et al., 2020). In connection with this, Chapter 4 also provided animated reconstructions of driver-pedestrian interactions including gaze. Such visualizations may find use in safety systems that enhance the situational awareness of road users.

Despite the controlled setting, the study obtained valuable insights into the nature of eye contact on the road. The observed (objective) eye contact durations were approximately 0.8–1 s for standing pedestrians and approximately 2.4 s for crossing pedestrians in their interactions with a driver. Although these durations were likely different from those in real traffic due to the externally imposed conditions of the experiment and the presence of other salient stimuli on a real road, they illustrated the time-critical and dynamic nature of driver-pedestrian gaze interactions.

There were also some limitations to Chapter 4. First, the proposed operationalization of driver-pedestrian eye contact did not consider the subjective awareness of the persons involved. This is a shortcoming of eye-tracking technology in general, i.e., being susceptible to the "looking but not seeing" phenomenon, where gazing at a

target does not necessarily imply that the person is aware of it or cognitively processing it (White & Caird, 2010). Another assumption in the study was that if both the driver and the pedestrian were looking at each other's faces, they were looking at each other's eyes in an attempt to make eye contact. At the distances used in the experiment, and given the video resolutions and accuracies of the eye-trackers, it was not possible to reliably distinguish gaze on individual facial features.

Next, while parts of the eye contact detection method were automated, e.g., in the computer vision to determine the locations of the driver and the pedestrian, other aspects such eye-tracker synchronization and data processing remained largely manual tasks. Further, the algorithm did not function in real-time, which would be necessary before an eye contact detection system could be deployed on the road. Nonetheless, its design with frame-by-frame processing in mind showed potential for adaptation into real-time applications in the future, via a combination of custom software and powerful yet portable hardware to process the data streams.

The artificial nature of the experimental setup, involving a stationary vehicle, pedestrians equipped with expensive mobile eye-trackers, and staged indoor interactions, also limited the generalizability of the findings. Finally, the custom and rather rudimentary computer vision techniques used for vehicle and driver location estimation, left room for improvement. While this was sufficient in Chapter 4's highly controlled setting, more sophisticated object detection algorithms would be required in more complex, real-world traffic environments and showed promise in preliminary tests. It is worth noting that computer vision techniques, e.g., YOLO, have considerably improved in terms of inference speed and accuracy since the publication of Chapter 4. It is expected that in the near future, real-time and automatic object detection in traffic combined with eye-tracking will become the norm.

Future research could incorporate methods such as think-aloud protocols or event recorders alongside eye-tracking to bridge the aforementioned gap between objective and subjective measures of eye contact. Real-time operation, powered by latest advances in object detection algorithms and computing power, would be another goal to strive for. Future studies could also consider naturalistic and outdoor scenarios to obtain more generalizable results, and take advantage of advancements in mobile eye-tracking technology (e.g., Khaldi et al., 2020; Robert et al., 2024) to overcome problems such as infrared interference due to sunlight. Road user eye contact detection is also relevant to the first five SAE levels of driving, where the attention of the 'driver' can be intermittent or absent. The missing natural eye contact in traffic interactions with such vehicles might create a need for artificial substitutes, such as (anthropomorphic) external human-machine interfaces (eHMIs) to communicate between the vehicle and VRUs (e.g., Jaguar Land Rover, 2018). Eye contact detections could also serve as an input for wearable devices equipped

with safety systems, providing warnings to drivers and pedestrians or feeding into AV control modules.

Again, this paper was accompanied by open-sourced data and code for the sake of scientific transparency and replication. Chapter 4 served as an early step towards real-time object and eye contact detection of road users in combination with eye-tracking glasses, and the automated analysis of mobile eye-tracking data. Chapter 5 built on this foundation to create concepts of safety systems for road users.

6.4. Chapter 5

Chapter 5 presented four new concepts in the field of mobile eye-tracking and contextual risk assessment in traffic situations, using combinations of Tobii Pro Glasses 2 eye-tracking, YOLOv8 object detection, and the GPT-4V vision-language model (VLM). The main goal was to create concepts of a safety system capable of understanding traffic situations, assessing risk, and providing context-specific feedback to road users (in real-time, if possible). The working of these concepts were demonstrated via recordings made in four different environments: (1) eye-tracking video of an indoor dining table scene with various everyday objects, (2) eye-tracking video of a pedestrian navigating a parking garage, (3) dashcam video from a car driving on urban streets, and (4) eye-tracking video of a distracted pedestrian using a mobile phone in a street-crossing scenario. In the process, methods to automatically analyze and contextualize mobile eye-tracking video and gaze data were also developed, providing one solution to the long-standing problem of manual annotation in head-mounted eye-tracker studies (Franchak et al., 2018; Vabalas & Freeth, 2016).

The first concept system demonstrated that the integration of Tobii 2 and YOLOv8 via a Python script was feasible in near real-time at 25 fps to identify objects in view and if and when the user's gaze fell on them. The automatic identification and localization of the gaze target was achieved by checking (for every video frame) whether the user's instantaneous gaze point lay inside the dimensions of a detected (and classified) object's YOLOv8 bounding box, and if so, subsequently highlighting that bounding box. This concept showed potential as a tool to monitor the attention and situational awareness of (road) users and was tested in both the indoor and the parking garage scenarios.

Concept 2 demonstrated the potential to use GPT-4V to assess risk in urban traffic situations via a frame-by-frame analysis of a dashcam video from a car driving in Amsterdam (Young Niles, 2017). It was observed that GPT-4V's risk scores increased on busy streets and reduced on empty roads. It was also found that these risk ratings correlated strongly (r = 0.71) with human-provided risk ratings (Bazilinskyy et al., 2020), although there were a few instances of deviation between the two sets of scores. This was attributed to GPT-4V's inability to reliably interpret

details or specific objects in scenes, and its lack of understanding of temporal and spatial relationships between frames (Guan et al., 2024; Wen et al., 2024; Zhou & Knoll, 2024). Overall, the results indicated that GPT-4V had a good and intuitive (almost human-like) understanding of traffic risks based on visual inputs.

Concept 3 tested GPT-4V's ability to assess risk in the aforementioned pedestrian in a parking garage scenario from an eye-tracking video (De Winter et al., 2021). This video contained a gaze marker denoting the user's instantaneous gaze point but did not contain YOLOv8 object bounding boxes. The effect of the gaze marker on GPT-4V's perception of risk was found to be minimal, with a very strong correlation (r = 0.98) between risk ratings from two prompts, one asking GPT-4V to consider the gaze marker in its analysis and the other asking GPT-4V to ignore the gaze marker. The effect of frame resolution was also found to be minimal, with another very strong correlation (r = 0.97) between risk ratings for frames evaluated under the low-res and high-res modes of GPT-4V. Qualitative inspections of the GPT-4V outputs revealed that the model could not always correctly identify the user's gaze target, e.g., other pedestrians. Subsequent tests of GPT-4V's ability to identify any pedestrians lying behind the gaze marker produced a moderate correlation (r = 0.6) with a post-hoc YOLOv8 benchmark of pedestrian detections. GPT-4V also had a true-positive rate of 76% and a false-positive rate of 6.1% at detecting when a pedestrian was being looked at. Thus, in addition to the occasional failure in correctly identifying the gaze target, as noted earlier, GPT-4V sometimes also hallucinated, i.e., reported pedestrians under the gaze marker when there were none. These observations suggested that GPT-4V assessed frames (and by extension, risk) holistically but faced difficulties in pinpointing specific details within those frames, a limitation that could potentially be overcome by combining it with a second detection layer capable of reliably identifying individual objects, e.g., YOLOv8, and complementing the strengths of the two technologies. Overall, the results showed promise for using GPT-4V (or its successor) in a future safety system geared towards road users.

The fourth and final concept tested if GPT-4V and YOLOv8 could work effectively in tandem to assess risk in traffic. Frames from a staged eye-tracking video of a distracted pedestrian using a mobile phone and about to cross a busy street were analyzed by GPT-4V via four different prompts. The frames all contained the pedestrian's gaze marker but they were analyzed once with and once without object bounding boxes overlaid. The results for the different types of frames were observed to be in the following order of increasing risk: (1) empty road and no mobile phone use, (2) empty road and mobile phone use, (3) cyclists passing by and mobile phone use, (4) car passing by and mobile phone use, and (5) truck passing by and mobile phone use. This order was noted as having a high degree of face validity. Again, as also noted previously in Concept 3, the effect of considering or ignoring the gaze marker on risk ratings was minimal. However, explicitly stating the gaze target as a mobile phone in a prompt led to a significant increase in risk scores, compared to only stating that the pedestrian's gaze point was represented by a red circle (even

when the red circle fell on the mobile phone). This hinted that GPT-4V, by itself, could not actually 'see' the gaze marker or the object it fell on, and by extension, truly grasp their meaning and significance in the given context, and instead factored the text information about the gaze target in the prompts into its risk assessments. This idea mirrored earlier findings in Concepts 2 and 3 about GPT-4V's difficulty in individual features in a scene. The addition of object bounding boxes to the frames showed mixed, but overall promising results. Prompts with the gaze target stated explicitly in them did not benefit from an increase in risk scores due to the addition of bounding boxes, but prompts that only stated the function of the gaze marker did. This showed potential for using Concept 4 (and its combination of mobile eye-tracking, traditional computer vision, and generative AI) in real world traffic applications, as the gaze targets of road users are often not known in advance and therefore cannot be included in prompts.

Some limitations were identified in Chapter 5, particularly concerning real-time operation. While YOLOv8 had a latency of less than one eye-tracker video frame (< 40 ms), the eye-tracker itself (Tobii 2) had an inherent delay of 500–1000 ms in streaming video and gaze, which was too high for real-time functionality. The newer Tobii 3 eye-tracker was also evaluated and found to have an inherent delay of approximately 200 ms. Furthermore, GPT-4V took approximately 4 seconds on average to analyze a single prompt, which was too slow for dynamic and time-critical applications such as traffic safety. Clearly, new technological advancements in mobile eye-tracking and generative AI are necessary before real-time road user safety systems can be deployed.

Another important limitation was the high cost of putting together and operating such a concept safety system. Apart from mobile eye-trackers being prohibitively expensive, the running costs (and processing time) for analyzing several video frames multiple times each using GPT-4V also proved high, limiting the scope of the research. Future safety systems might instead benefit from using smaller, fine-tuned AI models that run directly on the mobile eye-tracker and which sample and analyze fewer, selected frames at lowered resolutions. The accuracy and reliability of GPT-4V's image assessments would also need more validation, especially with regard to spatial and temporal awareness, and GPT-4V's holistic approach, which proved prone to overlooking details and individual features.

Future research and development could focus on reducing latency and cost while improving accuracy and reliability. Potential applications to road safety might be advanced driver assistance systems (ADAS), wearable devices to improve the situational awareness of road users, and learner driver monitoring systems. Applications of such safety systems could also extend beyond the traffic environment to extreme sports, construction work, factory work, and search-and-rescue operations. Overall, the combination of Tobii 2 eye-tracking glasses, YOLOv8 object

detection, and GPT-4V image analysis was found to be a step in the right direction towards context-aware safety systems.

6.5. General discussion & conclusion

The results and limitations of the individual chapters recapped above offer scope for synthesis into more general findings regarding driver-pedestrian eye contact and the use of mobile eye-tracking in (future) traffic.

This thesis demonstrates that drivers' eye contact and its timing influence pedestrians' feeling of safety and crossing decisions (Chapter 2). It shows that eye contact in traffic is a dynamic phenomenon, and goes beyond previous research where eye contact was operationalized as a simple 'yes' or 'no' for a whole interaction (Chapters 2, 4). It also confirms the finding of earlier research that vehicle kinematics or implicit communication is a more dominant cue for pedestrians than eye contact (AlAdawy et al., 2019; Dey & Terken, 2017; Moore et al., 2019). With regard to this, the thesis shows that pedestrians can (and will) cross irrespective of drivers' eye contact, by relying on vehicle motion cues alone. Also, it was observed that in situations involving conflicting cues, i.e., mixed signals from a driver's eye contact vs. the vehicle's (lack of) movement, pedestrians prioritize the latter when making their crossing decision. The above findings do not mean that eye contact is unnecessary in driver-pedestrian interactions. This thesis proves that drivers' eye contact can provide clarity and be beneficial to pedestrians' perceived safety, especially at short distances and in situations where vehicle kinematics are ambiguous (see also Dey et al., 2019). Eye contact can also help reinforce the effect of vehicle motion cues on pedestrians, leading to a stronger, compounded effect. This occurs when both cues align intuitively to the pedestrian, e.g., the initiation or termination of the driver's eye contact occurs alongside braking (before yielding) or take-off (after yielding), respectively.

Similarly, on the other side of the road interaction, literature has noted that drivers most often use pedestrians' eye contact to decipher the crossing intentions of the latter in ambiguous situations (Schneemann & Gohl, 2016). Pedestrians' eye contact also makes drivers brake earlier, approach more slowly, and stop more often (Morgan et al., 1975; Ren et al., 2016; Snyder et al., 1974). This too, leads to increased perceived safety in both road users (Sucha et al., 2017) and lowers the risk of fatal collisions with pedestrians (Hussain et al., 2019; Richards, 2010; Tefft, 2013).

One reason why eye contact appears to have a similar type of effect on both drivers and pedestrians might be because eye contact has been shown to encourage compliance with rules, written or unwritten (Hamlet et al., 1984; Kleinke, 1980). It leads to the 'watching eyes' phenomenon, i.e., an impression in the mind of the person that they are being watched and that a certain (type of) action is expected of them, typically encouraging prosocial behavior and inhibiting antisocial behavior

(Sueur et al., 2023). In the case of a pedestrian on a curb 'receiving' (for lack of a better term) eye contact from an approaching driver, this likely manifests as a secondary cue to the pedestrian to cross (on top of the primary cue that is the reducing vehicle speed). In the case of an approaching driver 'receiving' eye contact from a pedestrian on a curb, this likely manifests as a secondary cue to the driver to slow down or stop (on top of the primary cue that is mere presence of the pedestrian at the curb).

Given the above results of this thesis, some key follow-up questions arise:

Q1. Are the added safety benefits of eye contact in traffic significant in practice?
Q2. Do they make driver-pedestrian eye contact worth investigating further?
Q3. Do they warrant the development of artificial substitutes, so that 'eye contact', in some altered form, can continue to exist between driverless, automated vehicles and pedestrians?

It is my opinion that driver-pedestrian interactions should not just be collision-free but also an efficient, comfortable, and pleasant experience for the road users involved. Eye contact is one of the most commonly used forms of non-verbal or explicit communication between drivers and pedestrians (Lee et al., 2021; Rasouli et al., 2017; Sucha et al., 2017). A sizable portion of pedestrians also report that eye contact is important for them to feel safe or that they are hesitant to cross the road in front of a vehicle in the absence of eye contact (Chapter 2). These feelings might stem from confirmations or self-reassurances that the driver is attentive, that they have (likely) been seen by the driver, or that they have done their duty to signal their intentions via eye contact before crossing, thereby placing their trust and the onus of safety on the driver. On the other hand, drivers might sometimes purposely avoid eye contact with pedestrians to assert their intention to not yield and claim right of way. Thus, it may be said that the action of eye contact (or lack thereof) is often accompanied by an unspoken transfer of the responsibility of safety in the interaction from one road user onto the other. Where initially this responsibility might have been borne equally by both road users, upon negotiation via (an avoidance of) eye contact, it shifts to fall largely on one party, who then modifies their behavior accordingly.

It may not even matter that many pedestrians are often unable to discern the eyes of drivers due to visibility challenges such as windshield glare, nighttime darkness, or poor cabin illumination (AlAdawy et al., 2019). What may matter more is that pedestrians (and possibly also drivers) *think* they have made eye contact (while having only looked at the silhouette of the other person and not actually discerned their eyes or registered the mutual gaze). In other words, performing the actions/motions of eye contact, e.g., head turning and gazing at the other road user, might be more important than mentally processing and acknowledging its occurrence. This *illusion* of eye contact alone might be sufficient to generate any

desired feelings of safety and help drive interactions forward smoothly. Broadly speaking, it may be more important to pedestrians to feel in a general sense that they have been noticed by drivers or AVs, than needing to explicitly see and acknowledge a pair of eyes looking back at them. This observation might be useful to narrow down the set of concept designs for eye contact substitutes in AV-pedestrian interactions and avoid over-engineering a solution.

Of course, the alternate approach would be to simply not engineer a replacement for eye contact in AVs and trust that road users will, in time, get accustomed to not having this cue. This is the simplest and most cost-effective route, but sacrifices the human/human-like element of interactions that offers a reassurance of safety. Thus, while it is likely that pedestrians will be able to adapt their crossing behavior to rely solely on vehicle motion cues in order to negotiate interactions with AVs, this would not be ideal in terms of comfort and pleasure. Therefore, it may be said that in terms of the MoSCoW prioritization method (Agile Business Consortium, 2014), the illusion of eye contact, or by extension, a well-chosen eye contact substitute, is a "should-have" in traffic interactions. A well-designed replacement for eye contact in AV-pedestrian interactions might also have the added benefit of simultaneously substituting other missing forms of driver non-verbal communication too, e.g., hand gestures and nodding. So, to answer the questions above:

A1. The extra benefits brought by eye contact in traffic concern perceived safety in the minds of drivers and pedestrians, and are significant in practice.
A2. Driver-pedestrian eye contact *is* worth studying further, especially in naturalistic settings, to maximize the aforementioned safety benefits and improve the quality of road interactions.
A3. Instead of artificial substitutes that imitate eye contact, e.g., anthropomorphic eyes on AVs (Chang et al., 2017; Jaguar Land Rover, 2018), it may be more effective to maintain/increase perceived safety levels via alternate cues (thereby rendering (the illusion of) eye contact superfluous), e.g., textual, icon, or light pattern eHMIs, to ensure the continuation of any added safety benefits in AV-pedestrian interactions. Moreover, this choice might also help avoid any 'uncanny valley' effects associated with the former approach, that may even prove detrimental to perceived safety.

Now, as discussed above, if it is the actions/motions of eye contact and its timing in traffic that really matter, then there is incentive to study and detect this phenomenon as accurately as possible. Therefore, this thesis evaluates, and subsequently makes use of, one of the most versatile and reliable gaze measurement techniques available, viz. (mobile) eye-tracking to objectively operationalize driver-pedestrian eye contact (Chapters 3, 4). With regard to the performance of Tobii mobile eye-trackers, the thesis found that accuracy was worse for targets at an eccentricity, but not necessarily for dynamic conditions, e.g., head turning or walking. Thus, the use of chinrests with mobile eye-trackers, e.g., in staged or simulated driver

eye-tracking studies, is not advisable because any potential improvements in accuracy from reduced head movement are insufficient to offset losses in accuracy from gazing at targets eccentrically. The newer Tobii 3 eye-tracker also consistently outperformed the older Tobii 2 model in terms of accuracy, especially with eccentric targets and in dynamic, walking conditions. It is therefore the better choice for future traffic studies, e.g., of naturalistic driver-pedestrian gaze interactions or individual road user gaze behavior.

However, mobile eye-tracking accuracies were found to be poorer than the manufacturer's specifications, which is in line with previous research (Niehorster et al., 2020). This might serve as a reminder to traffic researchers employing eye-tracking to exercise caution when taking published reports by manufacturers at face value, especially since the latter may have a vested interest in reporting favorable outcomes (Ioannidis, 2005). Open data will therefore be crucial to ensure that other researchers can reproduce and replicate findings in eye-tracking, traffic research, and beyond. In this thesis, all chapters are accompanied by open data and source code for the sake of transparency and to contribute unconditionally to the body of scientific knowledge.

With regard to the operationalization of eye contact in traffic, this thesis objectively and bi-directionally detects it for the first time, using dual, synchronized eye-tracking of a driver and a pedestrian, combined with computer vision techniques to estimate their locations relative to each other (Chapter 4). Partial automation of the eye-tracking analysis, with future potential for real-time applications, was a useful by-product of this endeavor, offering a viable alternative to the traditional approach to eye-tracking analysis that is manual and labor-intensive.

This thesis' definition of driver-pedestrian eye contact as mutual gaze within 4° of each road user's eyes, helps to classify interactions as either involving the cue or not. Geometrically, the 4° is the half-angle of an imaginary cone of gaze/eye contact, whose base lies around the eyes of one road user, e.g., the driver, and whose apex lies at the eyes of the other road user, e.g., the pedestrian (see Gamer & Hecht, 2007). The (apex) angle of this cone would therefore be 8°, demarcating the limits of gaze that could be considered eye contact, i.e., by the circumference of the base of the cone. A similar cone exists simultaneously in the opposite direction, i.e., with its apex at the driver's eyes and its base at the pedestrian's. Such precision in detecting eye contact would be useful for future safety systems to differentiate the former from gazes in the general directions of road users i.e., to ascertain whether an eye contact cue was given and intended to negotiate a road interaction. That said, it is worth remembering that this 4° threshold for eye contact was determined via a staged driver-pedestrian interaction at a representative distance; it would need to be validated across a broader range of naturalistic traffic scenarios and interaction distances before it could be applied in practice.

It is also worth noting that this 4° threshold for eye contact includes any inaccuracies in the eye-tracker itself. For example, the eye-tracking accuracy was as poor as 3.5° for pedestrians in a walking scenario (see Chapter 3). This, in turn, implies that the gaze of a walking pedestrian on a driver is likely quite accurate (i.e., 0.5°, assuming an additive relationship between human error and eye-tracker accuracy). In other words, pedestrians in Chapter 4 made eye contact with the driver by bringing him to the center of their foveal vision, and not by using their parafoveal or peripheral vision.

As mentioned earlier, such an objective operationalization cannot determine whether road users are consciously aware of eye contact. However, since an illusion of eye contact may be what ultimately matters, the potential lack of awareness is internal and likely irrelevant, since only external behavior is observable. In other words, any future sensor or practical application will rely on precisely detected eye movements and not cognitive models. In future studies, it might be interesting to use a think-aloud method or event recorders to examine what drivers and pedestrians experience moment-to-moment in an interaction (e.g., “I feel like I am being watched”, “The other person is making eye contact with me”, “I feel safe”), and to relate these introspections to their gaze directions and head orientations.

Objectivity of gaze measurements could potentially aid future ADAS, smart wearable devices, and AV control modules in maintaining road safety. In the future, an AV might be able to directly track or indirectly obtain the gaze directions of a pedestrian and its own “driver”/occupant, and factor events such as eye contact when deciding its kinematic and eHMI display behaviors. Eye-tracking performance would be critical in such scenarios, as false positives or false negatives when detecting eye contact could lead to confusion in the AV due to misalignment with other cues, e.g., ego vehicle/pedestrian kinematics, which in turn could lead to collisions or near crashes. Finally, it might be possible in the future that a network of connected AVs and AI-powered wearable devices, all capable of mobile/remote eye-tracking, could intercommunicate and exchange detections of road user gaze and other non-verbal cues, for optimal negotiation of interactions and road safety. Such a network might also include other vulnerable road users such as cyclists and motorcyclists, but will likely first have to contend with the challenges of mixed traffic in the immediate future.

In connection with the aforementioned wearable devices for future road (user) safety, this thesis presents four interrelated concepts of such systems that use a mixture of mobile eye-tracking, object detection, and AI (Chapter 5). These concepts arose from logically extending the idea of automatically detecting the target of a driver’s/pedestrian’s gaze and eye contact (Chapter 4), to the more broad application of real-time detection of a road user’s/vehicle’s view, the gaze target (if any), and the context of the traffic scene, to provide risk assessments and feedback to improve traffic safety. For example, the feedback might be regarding whether or not a

pedestrian noticed an oncoming car before crossing a street, or regarding which moments of driving/cycling in an urban environment pose a collision risk. Such feedback in real-time could prompt behavioral changes in road users such as reducing speed, stopping, or communicating via non-verbal means, e.g., visual scanning, eye contact, gestures, or eHMI display changes. In other situations, feedback might not be required, e.g., when the traffic environment as a whole is not dangerous, there are no hazards in the vicinity, or the road user is already attentive, in which cases, feedback should not be provided as it, by itself, might be a distraction.

The four concepts help take the frontiers of traffic research forward from post-hoc eye-tracking and context analyses of staged interactions to (partially) real-time analyses of live interactions, while identifying bottlenecks, e.g., GPT-4V inference times that, if resolved, would open the doors to truly real-time, deployable, and wearable safety systems in traffic. This thesis shows that GPT-4V's assessments of traffic risk already strongly resemble those of humans, although there is still room for the former's improvement in terms of hallucinations, identifying details, and understanding the influence of 3D space and time in traffic interactions. Combining GPT-4V with YOLOv8 helped mitigate some of these drawbacks, showing promise for future safety systems built into wearables. The goal of real-time operation may even be feasible within the next decade, given the rapid advancements in head-mounted eye-tracking, object detection, and AI. One promising example of this progress is the launch of the Apple Vision Pro mixed-reality headset, capable of eye-tracking (Apple Inc., 2024). Another is the fact that GPT has only been available since the last 2 years, and can already handle text, code, tabular data, documents, audio, images, and video (OpenAI, 2024).

Given the exponential trend of technological progress (Kurzweil, 2024), it is conceivable that in a couple of decades, vehicles (and machines in general) will be able to accurately assess the state of persons in their vicinity, e.g., drivers, pedestrians, factory workers, just as humans do with each other. These developments will have major implications for how the world is shaped, in the automotive domain and beyond. For instance, traffic interactions may change from human-human negotiations to human-machine and machine-machine negotiations, with pedestrians merely following action prompts, e.g., "Cross now" or "Don't cross", from smart traffic lights, XR headsets, or eHMI displays, and AV passengers being notified about interactions over which they have no direct control, e.g., via in-vehicle displays and speakers. In other fields, such as manufacturing, construction, surgery, and search-and-rescue, humans may assume entirely supervisory roles, and occasionally train or direct artificially intelligent machines using their gaze and eye-tracking. In the fields of education, psychotherapy, marketing, entertainment, professional sports, social interaction, and human-computer interaction, eye-tracking and AI might be used to understand the needs and preferences of people and tailor solutions and products accordingly. That said, affordability will also be necessary for

widespread adoption, since the technologies employed and their operating costs are currently too expensive for daily use by the average (road) user.

Finally, based on the results of the individual papers in this thesis, the general discussion, and recent trends in technology, some recommendations for future research into eye contact and eye-tracking are listed as follows:

1. Investigating the effect of (the timing of) pedestrians' eye contact on drivers, as a complementary study to Chapter 2.

2. Combining eye-tracking of road users with think-aloud protocols and event recorders to supplement objective detections of gaze with subjective reports of the awareness of gaze. This would help gain a better understanding of the psychological significance of eye contact in traffic, to supplement the more objective approach used in this thesis.

3. Detecting driver-pedestrian eye contact using dual eye-tracking in naturalistic conditions. This would be the logical next step from the staged scenarios in this thesis.

4. Validating the gaze threshold for driver-pedestrian eye contact across a range of interaction scenarios, distances, and configurations. This would supplement the finding of the 4° threshold in this thesis and form an important step forward towards practical applications.

5. Developing fully automated and real-time eye contact detection systems that can be deployed in traffic to improve safety.

6. Inventing methods to remotely and accurately detect the gaze and eye contact of pedestrians from a vehicle, thereby freeing them of the need to wear an eye-tracker.

7. Developing real-time assistance systems built into wearable devices that can analyze user gaze and the surroundings to provide appropriate feedback in a variety of tasks and environments. This idea may advance human safety and productivity and be a step towards human-machine symbiosis, but ethical and privacy concerns will likely arise.

8. Creating eye-tracking controlled and artificially intelligent systems that allow users to interact with them, perform complex virtual and physical actions, or assume a supervisory role using just their gaze.

In conclusion, this thesis provides insight into the importance and functional relevance of eye contact in traffic, benchmarks the accuracy of tools required to

objectively detect eye contact, operationalizes driver-pedestrian eye contact, presents prototypes of a real-time safety system that uses eye-tracking, object detection, and AI to provide feedback to road users, synthesizes the findings of the above four papers, offers recommendations for further research, and reflects on the future of mobile eye-tracking and AI.

6.6. Epilogue

Some of the findings of this thesis show that driver-pedestrian eye contact is a complex and dynamic phenomenon and not just a simple binary cue. In a recent, related study, in which I was not the first author, eye contact between drivers was explored to further the understanding of eye contact in traffic. The study found that driver-driver eye contact is employed (and avoided) for a variety of reasons, both prosocial and antisocial.

In short, an online survey was conducted with 3,857 respondents spread across 20 countries. Participants were shown an image of a busy roundabout and asked about their eye contact behavior (if any) and its likelihood in such a scenario. A free-response item recorded their rationale behind their responses. The answers of 600 respondents (199 from Mexico, 200 from the United States, and 201 from the Netherlands) were subsequently annotated, with the annotator unaware of the respondents' countries of origin.

The results in Figure 6.1 show the percentages of respondents for 15 categories annotated from the answers. The precise definitions of the response categories are provided in Table 6.1 below. Figure 6.1 reveals that reasons for eye contact range from gathering information about the other person's state or making oneself known, to assertively asking for or even enforcing right of way.

It is also noteworthy that drivers sometimes purposely *avoid* eye contact. For example, participants reported they do this to better focus on traffic (consistent with the earlier notion that implicit communication is more important than eye contact). Some drivers also reported avoiding eye contact due to it being stressful or uncomfortable, or to avoid conflict. A particularly interesting category is the avoidance of eye contact to pretend that another competing driver has not been noticed. This manipulative tactic is but one strategy to claim right of way. Many of these reasons for making and avoiding eye contact were also mentioned in the general discussion earlier.

Figure 6.1 also shows that there are statistically significant cultural differences. In Mexico, drivers are more likely to use eye contact, or the avoidance of it, to achieve their goals, while in the United States and the Netherlands, traffic rules often suffice. This may be because traffic is more strictly regulated by formal rules in wealthier countries, and so, non-verbal communication is less used.

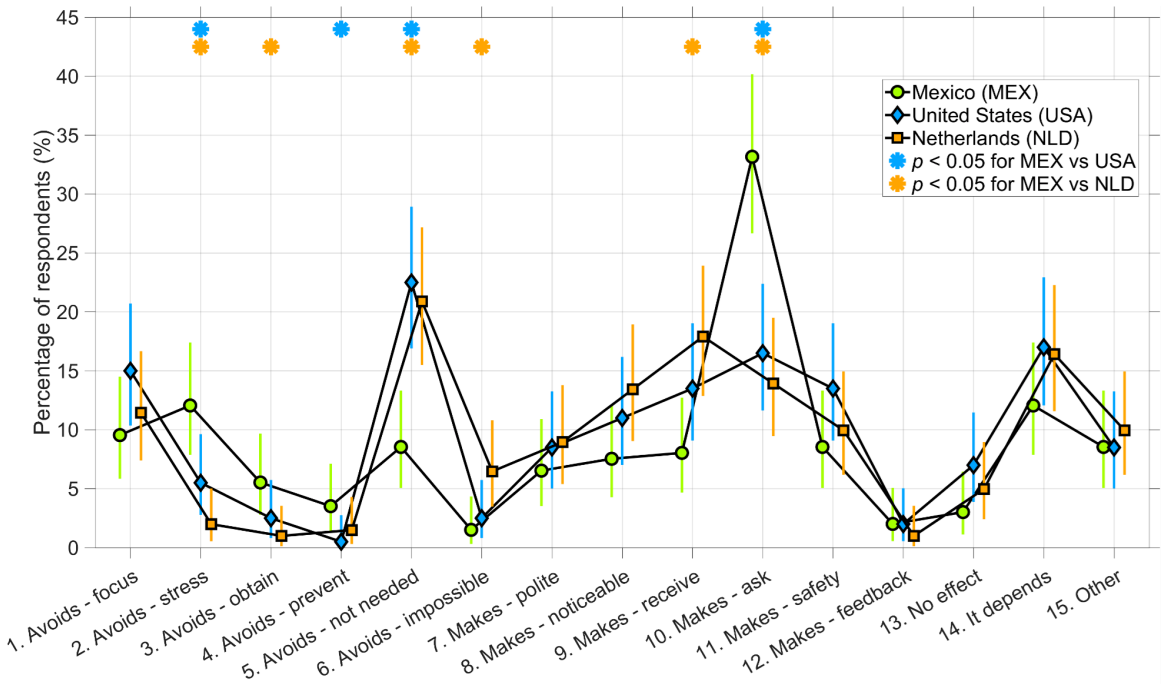

Figure 6.1. Percentage of responses categorized through a categorical content analysis for an open-ended question, in which respondents were asked to explain why they would make or avoid eye contact when driving in the given roundabout scenario. The figure includes 95% confidence intervals, and indicates statistical significance ($p < 0.05$) for comparisons between Mexico and the United States, and between Mexico and the Netherlands.

These findings suggest that automated vehicles will need a high level of social intelligence to operate effectively on today's roads, and eye-tracking and vision-language models could be useful to achieve this goal (cf. Chapter 5).

The current additional study also relates to earlier findings presented in Chapter 2 about drivers' eye contact having an impact on pedestrians' decisions when there is a crossing conflict. The present study also suggests that eye contact is more than just for gathering information, conveying presence, and making decisions; it can also be *actively used*, or *actively avoided*, to enforce outcomes in traffic.

In the end, it should be remembered that the experiments in both Chapters 2 and 5 were conducted in online environments. This offers strong potential for cross-cultural comparisons and standardization of measurements, but it remains somewhat arbitrary in nature and disconnected from real traffic. Further research into eye contact measurements in real-world settings, as in Chapters 4 (staged scenarios) and 5 (staged and naturalistic scenarios) will remain important.

Table 6.1
Categories into which the 600 responses to the open-ended question were annotated.

	Category	**Category**
1	Avoids - focus	Avoids eye contact in order to focus on the road / minimize distractions / out of safety considerations *Example: "Eye contact can be a distraction and can cause accidents."*
2	Avoids - stress	Avoids eye contact because it is stressful / uncomfortable / not liking eye contact in general Example: "*eye contact on its own is stressful...eye contact in a high stress environment like the described scenario, would drive me mad.*"
3	Avoids - obtain	Avoids eye contact to prevent yielding right of way / increase chance to enter Example: "*if you focus on your car and their car and don't make eye contact they are more likely to let you in*"
4	Avoids - prevent	Avoids eye contact to prevent irritating the other driver / getting intimidated Example: "*other drivers might be aggressive so I definitely avoid eye contact at any moment while driving.*"
5	Avoids - not needed	Does not use eye contact because is not needed; kinematics / blinkers / rules give the necessary information / it is safer to rely on objective information *Example: "I don't really make eye contact with other drivers in those situations. I just look at car's positions and make decisions based on that."*
6	Avoids - impossible	Does not make eye contact as it is not physically possible because of e.g., distance, speed, or window glare Example: "*Some vehicles also have darker tint than others, or drivers may be wearing sunglasses, so I might not be able to make eye contact.*"
7	Makes - politeness	Makes eye contact to achieve mutual awareness / out of politeness / reciprocation Example: "*I find more success on the road when I make eye contact with drivers to convey understanding and to be polite and considerate.*"
8	Makes - noticeable	Makes eye contact to indicate intentions / make oneself noticeable / indicate to the other drivers they have been noticed Example: "*I feel that making eye contact ensures the other driver sees you and know you are there*"
9	Makes - receive	Makes eye contact to receive information about the other driver's intentions or state / to see if the other driver has noticed them Example: "*might make eye contact to see what the other driver would do*'"
10	Makes - ask	Makes eye contact to increase the chance to get right of way / ask permission Example: "*giving eye contact in my opinion increases your chance of getting into the roundabout because other drivers can see that you're trying to enter.*"
11	Makes - safety	Makes eye contact for safety reasons Example: "*Eye contact is a good way to avoid an accident*"
12	Makes - feedback	Makes eye contact to thank or criticize an action already taken Example: "*I don't like to make eye contact I just do it when they do something wrong*"
13	No effect	Eye contact might not have an effect / can happen accidentally / not on purpose Example: "'*I don't feel that eye contact should change the flow of traffic*"
14	It depends	Makes or avoids eye contact depending on the situation or culture Example: "*Everything depends on the way the roundabout works and the place you are driving*"
15	Other	Non-specific / unclear / other Example: "*Eye contact is very important when you drive*"

The abstract of this paper is provided below:

Abstract

The advent of self-driving cars has sparked discussions about eye contact in traffic, particularly due to challenges automated vehicles face in non-verbal communication with human road users. In his 1992 book, Turn Signals Are The Facial Expressions Of Automobiles, Don Norman describes how drivers in Mexico City deliberately avoid eye contact when entering a roundabout to create uncertainty in the minds of other drivers, leading the latter to yield right of way. Norman also argued that such manipulative or aggressive behavior would not be tolerated in the United States. In the present study, we tested these claims through an online survey involving 3,857 respondents from 20 countries. The results confirmed that Mexican drivers reported a higher frequency of non-speeding 'aggressive' violations compared to those from most other countries. Regarding eye contact in roundabout scenarios, national differences were found not so much in the frequency of eye contact but in the reasons behind its use, or lack thereof. Mexican drivers tended to avoid eye contact to reduce tension or avoid conflict with other drivers. However, they also frequently reported making eye contact to assert or subtly enforce their right of way. In higher-income countries like the United States, driver-driver eye contact is often deemed unnecessary. In conclusion, our findings partially correspond with Norman's anecdote based on his experiences in 1950s Mexico City. These results may have implications for understanding the stability of traffic cultures and the challenges related to eye contact and non-verbal communication faced by developers of automated vehicles.

Citation:

De Winter, J. C. F., **Onkhar, V.**, & Dodou. (2025). Cross-national differences in drivers' eye contact and traffic violations: An online survey across 20 countries. *Transportation Research Part F: Traffic Psychology and Behaviour*, *109*, 711–725. https://doi.org/10.1016/j.trf.2024.12.021

References

Agile Business Consortium. (2014). Chapter 10: MoSCoW Prioritisation. https://www.agilebusiness.org/dsdm-project-framework/moscow-prioririsation.html

AlAdawy, D., Glazer, M., Terwilliger, J., Schmidt, H., Domeyer, J., Mehler, B., Reimer, B., & Fridman, L. (2019). Eye contact between pedestrians and drivers. *Proceedings of the Tenth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Santa Fe, New Mexico, 301–307. https://pubs.lib.uiowa.edu/driving/article/28343/galley/136635/view

Apple Inc. (2024). Apple Vision Pro. https://www.apple.com/apple-vision-pro

Bazilinskyy, P., Eisma, Y. B., Dodou, D., & De Winter, J. C. F. (2020). Risk perception: A study using dashcam videos and participants from different world regions. *Traffic Injury Prevention, 21*, 347–353. https://doi.org/10.1080/15389588.2020.1762871

Chang, C.-M., Toda, K., Sakamoto, D., & Igarashi, T. (2017). Eyes on a car: An interface design for communication between an autonomous car and a pedestrian. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Oldenburg, Germany, 65–73. https://doi.org/10.1145/3122986.3122989

De Winter, J., Bazilinskyy, P., Wesdorp, D., De Vlam, V., Hopmans, B., Visscher, J., & Dodou, D. (2021). How do pedestrians distribute their visual attention when walking through a parking garage? An eye-tracking study, *Ergonomics, 64*, 793–805. https://doi.org/10.1080/00140139.2020.1862310

Dey, D., & Terken, J. (2017). Pedestrian interaction with vehicles: Roles of explicit and implicit communication. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Oldenburg, Germany, 109–113. https://doi.org/10.1145/3122986.3123009

Dey, D., Walker, F., Martens, M., & Terken, J. (2019). Gaze patterns in pedestrian interaction with vehicles: Towards effective design of external human-machine interfaces for automated vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Utrecht, the Netherlands, 369–378. https://doi.org/10.1145/3342197.3344523

Faas, S. M., Stange, V., & Baumann, M. (2021). Self-driving vehicles and pedestrian interaction: Does an external human-machine interface mitigate the threat of a tinted windshield or a distracted driver? *International Journal of Human–Computer Interaction, 37*, 1364–1374. https://doi.org/10.1080/10447318.2021.1886483

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2018). See and be seen: Infant–caregiver social looking during locomotor free play. *Developmental Science, 21*, Article e12626. https://doi.org/10.1111/desc.12626

Gamer, M., & Hecht, H. (2007). Are you looking at me? Measuring the cone of gaze. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 705–715. https://doi.org/10.1037/0096-1523.33.3.705

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2024). HALLUSIONBENCH: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 14375–14385. https://doi.org/10.1109/CVPR52733.2024.01363

Hamlet, C. C., Axelrod, S., & Kuerschner, S. (1984). Eye contact as an antecedent to compliant behavior. *Journal of Applied Behavior Analysis, 17*, 553–557. https://doi.org/10.1901/jaba.1984.17-553

Holmqvist, K. (2017). *Common predictors of accuracy, precision and data loss in 12 eye-trackers.* ResearchGate. https://doi.org/10.13140/RG.2.2.16805.22246

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye-tracker hardware and its properties. In *Eye tracking: A comprehensive guide to methods and measures* (pp. 9–64). OUP.

Hussain, Q., Feng, H., Grzebieta, R., Brijs, T., & Olivier, J. (2019). The relationship between impact speed and the probability of pedestrian fatality during a

vehicle-pedestrian crash: A systematic review and meta-analysis. *Accident Analysis & Prevention*, *129*, 241–249. https://doi.org/10.1016/j.aap.2019.05.033

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*, Article e124. https://doi.org/10.1371/journal.pmed.0020124

Jaguar Land Rover. (2018). The virtual eyes have it. https://web.archive.org/web/20181203150349/https://www.jaguarlandrover.com/2018/virtual-eyes-have-it

Khaldi, A., Daniel, E., Massin, L., Kärnfelt, C., Ferranti, F., Lahuec, C., Seguin, F., Nourrit, V., & De Bougrenet de la Tocnaye, J. L. (2020). A laser emitting contact lens for eye tracking. *Scientific Reports*, *10*, Article 14804. https://doi.org/10.1038/s41598-020-71233-1

Kleinke, C. L. (1980). Interaction between gaze and legitimacy of request on compliance in a field setting. *Journal of Nonverbal Behavior, 5*, 3–12. https://doi.org/10.1007/bf00987050

Kurzweil, R. (2024). *The singularity is nearer: When we merge with AI.* Penguin Random House.

Lee, Y. M., Madigan, R., Giles, O., Garach-Morcillo, L., Markkula, G., Fox, C., Camara, F., Rothmueller, M., Vendelbo-Larsen, S. A., Holm Rasmussen, P., Dietrich, A., Nathanael, D., Portouli, V., Schieben, A., & Merat, N. (2021). Road users rarely use explicit communication when interacting in today's traffic: Implications for automated vehicles. *Cognition, Technology & Work, 23*, 367–380. https://doi.org/10.1007/s10111-020-00635-y

Lévêque, L., Ranchet, M., Deniel, J., Bornard, J.-C., & Bellet, T. (2020). Where do pedestrians look when crossing? A state of the art of the eye-tracking studies. *IEEE Access, 8*, 164833–163843. https://doi.org/10.1109/ACCESS.2020.3021208

Malmsten Lundgren, V., Habibovic, A., Andersson, J., Lagström, T., Nilsson, M., Sirkka, A., Fagerlönn, J., Fredriksson, R., Edgren, C., Krupenia, S., & Saluäär, D. (2017). Will there be new communication needs when introducing automated vehicles to the urban context? In N. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation. Advances in intelligent systems and computing* (pp. 485–497). Springer. https://doi.org/10.1007/978-3-319-41682-3_41

Moore, D., Currano, R., Strack, G. E., & Sirkin, D. (2019). The case for implicit external human-machine interfaces for autonomous vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Utrecht, the Netherlands, 295–307. https://doi.org/10.1145/3342197.3345320

Morgan, C. J., Lockard, J. S., Fahrenbruch, C. E., & Smith, J. L. (1975). Hitchhiking: Social signals at a distance. *Bulletin of the Psychonomic Society*, *5*, 459–461. https://doi.org/10.3758/BF03333299

Nathanael, D., Portouli, E., Papakostopoulos, V., Gkikas, K., & Amditis, A. (2019). Naturalistic observation of interactions between car drivers and pedestrians in high density urban settings. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International*

Ergonomics Association (pp. 389–397). Springer. https://doi.org/10.1007/978-3-319-96074-6_42

Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2018). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, *50*, 213–227. https://doi.org/10.3758/s13428-017-0863-0

Niehorster, D. C., Santini, T., Hessels, R. S., Hooge, I. T. C., Kasneci, E., & Nyström, M. (2020). The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods, 52*, 1140–1160. https://doi.org/10.3758/s13428-019-01307-0

Norman, D. A. (1992). *Turn signals are the facial expressions of automobiles.* Addison-Wesley Publishing Company.

Nyström, M., Andersson, R., Holmqvist, K., & Van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, *45*, 272–288. https://doi.org/10.3758/s13428-012-0247-4

OpenAI. (2024, August 8). GPT-4o system card. https://openai.com/index/gpt-4o-system-card

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Agreeing to cross: How drivers and pedestrians communicate. *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium*, Los Angeles, CA, 264–269. https://doi.org/10.1109/IVS.2017.7995730

Ren, Z., Jiang, X., & Wang, W. (2016). Analysis of the influence of pedestrians' eye contact on drivers' comfort boundary during the crossing conflict. *Procedia Engineering*, *137*, 399–406. https://doi.org/10.1016/j.proeng.2016.01.274

Richards, D. C. (2010). *Relationship between speed and risk of fatal injury: Pedestrians and car occupants* (Road Safety Web Publication No. 16). London, UK: Department of Transport.

Robert, F.-M., Otheguy, M., Nourrit, V., & De Bougrenet de la Tocnaye, J.-L. (2024). Potential of a laser pointer contact lens to improve the reliability of video-based eye-trackers in indoor and outdoor conditions. *Journal of Eye Movement Research*, *17*, Article 5. https://doi.org/10.16910/jemr.17.1.5

Rodríguez Palmeiro, A., Van der Kint, S., Vissers, L., Farah, H., De Winter, J. C. F., & Hagenzieker, M. (2018). Interaction between pedestrians and automated vehicles: A Wizard of Oz experiment. *Transportation Research Part F: Traffic Psychology and Behaviour, 58*, 1005–1020. https://doi.org/10.1016/j.trf.2018.07.020

Schneemann, F., & Gohl, I. (2016). Analyzing driver-pedestrian interaction at crosswalks: A contribution to autonomous driving in urban environments. *Proceedings of the 2016 IEEE Intelligent Vehicles Symposium*, Gothenburg, Sweden, 38–43. https://doi.org/10.1109/IVS.2016.7535361

Snyder, M., Grather, J., & Keller, K. (1974). Staring and compliance: A field experiment on hitchhiking. *Journal of Applied Social Psychology*, *4*, 165–170. https://doi.org/10.1111/j.1559-1816.1974.tb00666.x

Stuart, S., Alcock, L., Godfrey, A., Lord, S., Rochester, L., & Galna, B. (2016). Accuracy and re-test reliability of mobile eye-tracking in Parkinson's disease and

older adults. *Medical Engineering & Physics, 38*, 308–315. https://doi.org/10.1016/j.medengphy.2015.12.001

Sucha, M., Dostal, D., & Risser, R. (2017). Pedestrian-driver communication and decision strategies at marked crossings. *Accident Analysis & Prevention, 102*, 41–50. https://doi.org/10.1016/j.aap.2017.02.018

Sueur, C., Piermattéo, A., & Pelé, M. (2023). Eye image effect in the context of pedestrian safety: A French questionnaire study. *F1000Research, 11*, Article 218. https://doi.org/10.12688/f1000research.76062.2

SWOV. (2023). Road deaths in the Netherlands: SWOV fact sheet, September 2023. https://swov.nl/en/fact-sheet/road-deaths-netherlands

Tefft, B. C. (2013). Impact speed and a pedestrian's risk of severe injury or death. *Accident Analysis & Prevention*, *50*, 871–878. https://doi.org/10.1016/j.aap.2012.07.022

Tobii AB. (2017). Eye tracker data quality report: Accuracy, precision and detected gaze under optimal conditions—controlled environment, Tobii Pro Glasses 2.

Tobii AB. (2022). Tobii Pro Glasses 3 data quality test report: Accuracy, precision, and data loss under controlled environment (Rev. 1).

Vabalas, A., & Freeth, M. (2016). Brief report: Patterns of eye movements in face to face conversation are associated with autistic traits: Evidence from a student sample. *Journal of Autism and Developmental Disorders, 46*, 305–314. https://doi.org/10.1007/s10803-015-2546-y

Wen, L., Yang, X., Fu, D., Wang, X., Cai, P., Li, X., Ma, T., Li, Y., Xu, L., Shang, D., Zhu, Z., Sun, S., Bai, Y., Cai, X., Dou, M., Hu, S., Shi, B., & Qiao, Y. (2024). On the road with GPT-4V(ision): Explorations of utilizing visual-language model as autonomous driving agent. *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria. https://openreview.net/pdf?id=2UBexKm8TE

White, C. B., & Caird, J. K. (2010). The blind date: The effects of change blindness, passenger conversation and gender on looked-but-failed-to-see (LBFTS) errors. *Accident Analysis & Prevention*, *42*, 1822–1830. https://doi.org/10.1016/j.aap.2010.05.003

World Health Organization. (2023, December 13). Global status report on road safety 2023. https://www.who.int/publications/i/item/9789240086517

Yang, S. (2017). *Driver behavior impact on pedestrians' crossing experience in the conditionally autonomous driving context* (MSc thesis). KTH Royal Institute of Technology. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-220545

Young Niles. (2017, October 10). *Driving through Amsterdam with a dashcam* [Video]. YouTube. https://www.youtube.com/watch?v=CPWB9tZhT80&t=298s

Zhou, X., & Knoll, A. C. (2024). *GPT-4V as traffic assistant: An in-depth look at vision language model on complex traffic events.* arXiv. https://doi.org/10.48550/arXiv.2402.02205

Curriculum Vitae

11-03-1996	Born in Chennai, India

Education

1999–2013	Schooling Abacus Montessori School, Chennai, India
2013–2017	B. E. Mechanical Engineering SSN College of Engineering, Chennai, India
2017–2020	MSc Mechanical Engineering Delft University of Technology, Delft, The Netherlands
2020–2025	PhD “From sight to insight: Eye contact and eye-tracking in the driver-pedestrian context” Delft University of Technology, Delft, The Netherlands

Work Experience

2020–2024	Columnist TU Delta, Delft, The Netherlands
2020–2024	PhD Council Member Delft University of Technology, Delft, The Netherlands
2019–2020	Journalist Delft University of Technology, Delft, The Netherlands
2018	Human Factors Intern NEXTdriver, Delft, The Netherlands

List of Publications

Journal Publications

1. **Onkhar, V.**, Bazilinskyy, P., Stapel, J. C. J., Dodou, D., Gavrila, D., & De Winter, J. C. F. (2021). Towards the detection of driver–pedestrian eye contact. *Pervasive and Mobile Computing, 76*, Article 101455. https://doi.org/10.1016/j.pmcj.2021.101455 (Chapter 4 of this thesis)

   The data used was collected as part of the first author's MSc thesis:
   Onkhar, V. (2020). Algorithmic detection of eye contact in driver-pedestrian interactions.
   https://repository.tudelft.nl/islandora/object/uuid:b14b48ac-6df0-4a87-bb8d-5763b2697694.

   The journal article (Chapter 4) presents a complete rewrite and reanalysis of the MSc thesis work. Among other changes, the occurrence of eye contact was determined using computer vision instead of an inertial measurement unit (IMU), and gaze behavior was processed automatically instead of annotated manually. There is no textual overlap between Chapter 4 and the MSc thesis.

2. **Onkhar, V.**, Bazilinskyy, P., Dodou, D., & De Winter, J. C. F. (2022). The effect of drivers' eye contact on pedestrians' perceived safety. *Transportation Research Part F: Traffic Psychology and Behaviour, 84*, 194–210. https://doi.org/10.1016/j.trf.2021.10.017 (Chapter 2 of this thesis)

3. **Onkhar, V.**, Dodou, D., & De Winter, J. C. F. (2024). Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions. *Behavior Research Methods, 56*, 4221–4238. https://doi.org/10.3758/s13428-023-02173-7 (Chapter 3 of this thesis)

4. **Onkhar, V.**, Kumaaravelu, L. T., Dodou, D., & De Winter, J. C. F. (2024). *Towards context-aware road user safety systems: Design explorations using eye-tracking, object detection, and GPT-4V* [Manuscript submitted for publication]. (Chapter 5 of this thesis)

5. De Winter, J. C. F., **Onkhar, V.**, & Dodou, D. (2025). Cross-national differences in drivers' eye contact and traffic violations: An online survey across 20 countries. *Transportation Research Part F: Traffic Psychology and Behaviour*, *109*, 711–725. https://doi.org/10.1016/j.trf.2024.12.021
   (Part of this study is covered in the epilogue of this thesis)

Open Datasets

1. **Onkhar, V.**, Bazilinskyy, P., Stapel, J. C. J., Dodou, D., Gavrila, D., & De Winter, J. C. F. (2022). *Supplementary data for the paper 'Towards the detection of driver–pedestrian eye contact'* [Dataset]. 4TU.ResearchData. https://doi.org/10.4121/15134037

2. **Onkhar, V.**, Bazilinskyy, P., Dodou, D., & De Winter, J. C. F. (2022). *Supplementary data for the paper 'The effect of drivers' eye contact on pedestrians' perceived safety'* [Dataset]. 4TU.ResearchData. https://doi.org/10.4121/16866709

3. **Onkhar, V.**, Dodou, D., & De Winter, J. C. F. (2024). *Supplementary data for the paper 'Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions'* [Dataset]. 4TU.ResearchData. https://doi.org/10.4121/442018c6-30eb-4439-a452-c0046726905c

4. De Winter, J. C. F., **Onkhar, V.**, Dodou, D. (2025). *Supplementary data for the paper 'Cross-national differences in drivers' eye contact and traffic violations: An online survey across 20 countries'* [Dataset]. 4TU.ResearchData. https://doi.org/10.4121/94ddb48b-3a57-453e-861a-fab2da9f947b

Acknowledgements

The acknowledgements section is a difficult-to-write part of a thesis, since so many people contribute directly or indirectly to bring it to fruition that it is almost impossible to mention them all. This thesis is no different; my name might be the one on the front cover, but it is a placeholder for all those who helped shape this PhD. So, bear with me while I attempt to list everyone who influenced me over the last few years, and know that even if you are not here in writing, you are most certainly here in spirit.

First and foremost, I owe a massive debt of thanks to my supervisors, professors **Joost** and **Dimitra**, for their patience, understanding, and guidance throughout this long journey. Over the years, I have come to know them as not just gifted researchers, but also unfailingly kind people. On many an occasion, they also played the role of confidante, mentor, therapist, and career counselor, for which I am very grateful. They always went the extra mile with their supervision, and I could not have asked for better persons and scientists to watch and learn from.

On the subject of professors, I must thank some others at TU Delft, who, through their kindness and wisdom, were additional role models for me - **David**, **Marjan**, **Haneen**, **Riender**, and **Arkady**, although they may not know it. Every interaction with them has been a joy, and I have always come away with new knowledge and a smile on my face.

I would also like to extend my heartfelt thanks to those who contributed directly to the studies in this thesis. To **Pavlo** and **Jork**, my co-authors, former supervisors and friends, I am very grateful for your research guidance, keen life advice, and our many jolly conversations over dinner. To my other co-author, former student and friend, **Lokkesh**, thank you for all the wonderful evenings we spent together conducting experiments, exploring Indian restaurants, and traveling the Netherlands. I consider myself lucky to have supervised your MSc. Big thanks also to **Sparsh** and **Malvika** for their kindness and for always lending a hand, with experiments or otherwise. A word of appreciation to **Akash** and **Frederike** for their help as well. And finally, thank you **Lars**, for a great many things. Your ideas, advice, assistance with experiments, emotional support, memes, gaming sessions, and cat videos have all been stellar. You and **Tamika** may be far away in Australia, but you are both very near to my heart.

Next, I'd like to thank my dear friends, **Anand Sudha** and **Arjun**, for standing by me all these years since the days of our bachelor's degrees in India, keeping me grounded, always lending a friendly ear to my troubles, and offering sound advice

and moral support at every turn. If not for them, I might not even have come to the Netherlands, let alone completed a PhD. In the same breath, I'd like to express my gratitude to **Anand Krishnamurthy**, who I've known for just as long, and who has consistently proved a true friend in times of need. I am so glad you decided to follow me to the Netherlands. Of course, I must also acknowledge those who joined this friend group a little later, during our master's together, and who have remained close ever since - **Manav**, **Amogh**, **Akshar**, **Chaitanya**, **Sergin**, and others. Thanks for bringing some much-needed warmth and familiarity to what can otherwise be a cold and dark country for most of the year.

Speaking of cold and dark, **Siri**, who hails from Norway, was the first fellow PhD I had the good fortune to befriend at the start of this journey and the onset of the pandemic. What started as online bonding over cats and dry humor eventually grew into a close friendship cemented by mutual rants about life and work. I hope we keep this connection alive until we're both old and gray, because the PhD experience would have been dull without you. And where there is a Siri, there is usually also a **Wilbert** nearby. Wilbert and I became friends over our shared interest in fiction, art, and culture, but he soon proved to also have a gift for research, networking, and living life to the fullest, and so I've tried to take a leaf out of his book ever since. Here, I would be remiss if I did not mention **Xiaolin** and the other **SHAPE-IT members** (organizers and students alike) for giving me the opportunity to join the workshops and activities of their project as one of their own.

Special thanks also to my officemates, **Tom**, **Olger**, **Max**, **Italo**, **Nicky**, **Dennis**, and other past and present colleagues for all the academic discussions and fun conversations whenever I was (infrequently) in the office. Tom and Olger, in particular, have grown to become good friends of mine outside of work, via summer schools, drinks, dinners, and storytelling performances. I must recognize my colleagues-turned-friends from the department at large as well, such as **Alex**, **Ashwin**, **Linda**, and others, with whom I've enjoyed many nail-biting chess games and insightful conversations. A word of thanks also to the support staff of the department, especially **Hanneke**, who greatly facilitated paperwork throughout my PhD.

I also wouldn't have gotten far without the support of my Dutch friends, who graciously welcomed me into their homes and their circles all those years ago. **Rick**, **Luuk**, and I have come a long way since we first teamed up for course assignments during our master's. Through our regular dinners and nights out, we've stayed close ever since, in some cases, quite literally as neighbors! Via them, I've also had the pleasure to befriend their lovely network. On the subject of (former) neighbors, I'd like to thank **Timothy**, **Floris**, **Daan**, and their partners, whom I've watched grow alongside me from neighbors to classmates to good friends over the past seven years. I wouldn't trade our late-night philosophical discussions and fun excursions for anything in the world! Finally, I must convey my gratitude to **Maurice** and **Saskia** for

their love and encouragement that began over games of table tennis during our master's internships together and endures to this day. Looking forward to years more of long walks in nature, bookstores, mini-golf, and Indian cuisine!

I was also involved with the **PhD Council** of the Mechanical Engineering faculty for almost the entire duration of my PhD. I joined during the early days of the pandemic, initially as a way to keep informed about faculty/university news and events, but ended up staying for the wonderful people I met. Thank you **Mascha**, **Paul**, **Hans**, **Vasu**, **Suriya**, **José**, **Jette**, **Anna**, **Annabel**, **Pieter**, **Robin**, **Iris**, **Kate**, and the other past and present members, for over four years of great times organizing events, drinks, and dinners! I am really going to miss being part of such a vibrant team. Here, I'd also like to highlight **Hugo**, who has been like a brother to me ever since our days on the PhD Council, and a friendship with whom has seen each other through many ups and downs in life. I will visit you and **Sofia** in Italy very soon!

I also owe a debt of thanks to **Saskia Bonger** and the people at **Delta**, the newspaper of TU Delft, for giving me the opportunity (and the freedom) to write for them since the days of my master's. Composing monthly columns has been a welcome distraction to take my mind off research and to keep my critical thinking skills sharp, especially on topics outside science.

Thanks also to an assortment of friends and colleagues whom I know from across Mechanical Engineering, TU Delft, and the city of Delft at large, for their life advice and support - **Nienke**, **Jon**, **Tijl**, **Lot**, **Merel**, **Evelien**, **Lena**, **Tanya**, and **Johan**. I am also grateful to my landlord **Bart** for his kindness, friendship, and encouragement.

Lieve **Charlotte**, bedankt voor je grenzeloze liefde, geduld, zachtaardigheid, attentheid en zorg tijdens dit laatste deel van mijn PhD. Zonder jouw steun had ik dit niet kunnen doen. Ik prijs mezelf gelukkig dat ik een geweldige vrouw als jij heb ontmoet, en ik hou heel veel van je.

Last but not least, I'd like to thank my parents, **Kanthi** and **Narayan Onkar**, and my extended family, for their unwavering affection and encouragement. *Amma* and *Appa*'s emotional, physical, and financial support for my well-being, academics, and career have been constant over the decades. I would not be the person I am today without their love and everyday sacrifices, and I owe them more than I can repay. I dedicate this thesis to them.