Behavior Research Methods (2021) 53:430–446
https://doi.org/10.3758/s13428-020-01446-9

Does agreement mean accuracy? Evaluating glance annotation in naturalistic driving data

Reinier J. Jansen[1] · Sander T. van der Kint[1] · Frouke Hermens[1]

Published online: 29 July 2020

Abstract
Naturalistic driving studies often make use of cameras to monitor driver behavior. To analyze the resulting video images, human annotation is often adopted. These annotations then serve as the 'gold standard' to train and evaluate automated computer vision algorithms, even though it is uncertain how accurate human annotation is. In this study, we provide a first evaluation of glance direction annotation by comparing instructed, actual glance direction of truck drivers with annotated direction. Findings indicate that while for some locations high annotation accuracy is achieved, for most locations accuracy is well below 50%. Higher accuracy can be obtained by clustering these locations, but this also leads to reduced detail of the annotation, suggesting that decisions to use clustering should take into account the purpose of the annotation. The data also show that high agreement between annotators does not guarantee high accuracy. We argue that the accuracy of annotation needs to be verified experimentally more often.

Keywords Annotation · Glance perception · Accuracy · Reliability · Naturalistic driving

Introduction

Glance behavior by drivers plays an important role in safe driving, ranging from trajectory control (e.g., lane keeping, monitoring the distance to a lead vehicle) to higher-order skills (e.g., hazard perception, situation awareness), as well as involvement in secondary tasks (i.e., distraction). To improve traffic safety, it is therefore important to study how glance behavior is influenced by factors such as infrastructure, traffic conditions, non-driving tasks, and interactions with other road users.

'Naturalistic Driving' is a commonly used method to study glance behavior during everyday driving . Such studies often use human annotators to analyze video images from several inward and outward facing cameras in instrumented vehicles. In this study, we demonstrate how current practice of naturalistic driving glance annotation and analysis brings forward a methodological issue that has received relatively little attention, namely whether or not high agreement between annotators automatically means accurate annotation. We will illustrate this issue with a focus on analyzing the glance behavior of drivers.

Annotation of glance behavior

While manual annotation is often used to code human glance behavior, it is not the only means of recording people's glance direction. Eye trackers are an obvious alternative and often used in, for example, driving simulator studies, because they provide a real-time measure of glance direction, and because of their relatively high degree of reliability (if properly calibrated). However, eye trackers tend to perform worse in a real-life driving environment, where they can be affected by factors such as rapid changes in ambient illumination and vehicle vibration. Furthermore, high-end, accurate and precise eye trackers are expensive, limiting their large-scale use in instrumented vehicles (van Nes et al., 2019). Moreover, eye trackers, particularly those in the form of glasses (e.g., Tobii glasses, SMI glasses, Pupil labs eye tracker), may interfere with glance behavior while driving. Such interference can be due to the rim or frame of the glasses, that drivers may not yet be used to wearing, and the connecting cable that may

✉ Reinier J. Jansen
reinier.jansen@swov.nl

Sander T. van der Kint
sander.van.der.kint@swov.nl

Frouke Hermens
frouke.hermens@swov.nl

[1] SWOV Institute for Road Safety Research, P.O. Box 93113, 2509, AC The Hague, The Netherlands

influence the reliance on head rotation (possibly increasing the use of eye movements to shift gaze direction). As a consequence, the naturalistic aspect of drivers' driving behavior may be reduced. Some eye tracking systems, such as the system offered by Smart Eye, use cameras (in combination with infrared-light emitters) that can be mounted inside a vehicle, but as indicated, the cost of such systems prevents large scale use in naturalistic driving studies, limiting either the sample size, testing time, or both.

Studies of naturalistic driving have therefore strongly relied on manual annotation of images from cameras mounted inside vehicles, which are relatively low-cost compared to high-end eye trackers, such as the system from Smart Eye. The cameras minimally interfere with vision or glance behavior when mounted unobtrusively. To reduce the cost of manual annotation the focus of annotation is often on automatically detected events (e.g., harsh braking identified through an accelerometer), as opposed to processing all video data. Examples of uses of annotated glance behavior include studies of the impact of inattention on lane keeping performance (Peng et al., 2013), the impact of distraction from roadside objects (Belyusar et al., 2016), how the driving context and secondary task involvement influence glance behavior (Tivesten & Dozza, 2014), the effects of sensory cues on glance behavior prior to impending critical events in an ACC equipped car (Morando et al., 2016), crash causation factors (Dingus et al., 2006; Dingus et al., 2016), and differentiations between critical situations (Seppelt et al., 2017).

Annotators typically have several camera views at their disposal. For example, the UDRIVE project featured trucks equipped with three inward facing cameras (monitoring the driver's face, the cabin, and the driver's feet), and five outward facing cameras (monitoring the left blind spot, the front left, the front center, the front right, and the right blind spot). Information provided by one or more of these camera views may be used, e.g., to assess whether the driver is drowsy, to identify non-driving task involvement, and to identify in which direction a driver is looking. Manual annotation of glance direction is typically performed using a predefined set of glance locations, which appear to differ across studies. For example, Belyusar et al. (2016) studied the effect of digital billboards on naturalistic glance behavior in passenger cars by analyzing left forward and right forward glances. The number of these glances, as well as their durations, were found to increase in the presence of digital billboards, especially at the time the billboards transitioned between advertisements. Different regions were used by Seppelt et al. (2017), who clustered several glance locations into on-road glances (forward, left forward, right forward) and off-road glances (center stack, instrument cluster, interior object, cell phone, left mirror, left window, rearview mirror, right mirror, right window, passenger, eyes closed) to study which glance metrics could differentiate between crashes and near-crashes. It was found that a differentiation between crashes and near crashes could be made using the duration of on-road glances and the frequency of switches between on- and off-road locations. These examples show that annotated glance locations may be adapted to the research question of interest.

An important challenge in naturalistic driving studies is the large volume of data. For example, the SHRP2 database holds over 4300 years of naturalistic driving data collected with about 3400 car drivers (Hankey et al., 2016). Similarly, the UDRIVE database contains over 41,000 h of passenger car data and over 45,000 h of truck data (Van Nes et al., 2019). Frame-by-frame analysis of video data by human annotators is time-consuming, and costly to perform even when analysis focuses specifically on automatically detected events such as right-turns or abrupt braking, which are still sufficiently frequent in typical naturalistic driving studies to require extensive annotation work. Therefore, there has been a drive towards the development of computer algorithms to automatically annotate video images, for example, to predict where drivers are looking. For example, Fridman et al. (2016) developed a machine learning algorithm that classifies glance direction into six glance locations based on head pose (i.e., road, center stack, instrument cluster, rearview mirror, left, right). Importantly, the algorithm was trained and validated by means of video data from a field study, collected with a camera positioned on the dashboard and annotated by human coders. Likewise, Vora et al. (2017) trained their convolutional neural network using annotated naturalistic driving video data collected with a camera mounted near the rear-view mirror. Because of the use of human annotation as labels, the algorithms will work towards achieving the quality of this annotation, and therefore will be as good as the human annotation at best. Consequently, accurate human annotation is essential for the development of such algorithms.

Reliability versus accuracy

Naturalistic driving studies often use multiple annotators, who each code individual sections of the data to cope with the vast amount of data, and to increase reliability of the coding. The use of multiple annotators requires a method to evaluate and ensure that annotators use the same mental scheme for encoding. Two main approaches are used to establish the agreement between annotators and to deal with any disagreement between annotators. In one approach, all data or a selection is coded by at least two annotators. The joint annotations are then analyzed to get a sense of discrepancies. The extent of these discrepancies is often expressed in terms of one or more reliability measures, such as the percentage agreement or Krippendorff's alpha (e.g., Hayes & Krippendorff, 2007). Another approach circumvents the necessity of such measurements. If two annotators disagree in their coding, a third, mediating annotator is introduced to resolve the disagreement (e.g., Belyusar et al., 2016; Fridman et al., 2016), assuming

that the third annotator agrees with one of the two original annotators. In the latter case, the resulting percentage agreement is perfect by definition and inter-rater reliability is no longer a relevant concept.

In the above approaches it appears to be implicitly assumed that if two coders agree, their annotation will be correct. In line with this assumption, some studies explicitly refer to the annotated glance datasets as the 'ground truth' and subsequently use the annotated data to train and validate algorithms for automatic annotation (Tawari and Trivedi, 2014; Vora et al., 2017). Other studies seem to implicitly assume the annotation to be the ground truth, and then use the data for algorithm development (Belyusar et al., 2016; Fridman et al., 2016; Seppelt et al., 2017). It is important to realize that by using annotations as the ground truth, it is assumed that if two annotators agree, their annotation is correct. Whether this is actually the case has, however, rarely been studied, although some have raised the issue. For example, Naqvi et al. (2018) refrained from using a selection of existing glance datasets, because, amongst other reasons, "the information of ground-truth gaze position is not provided." (p. 15). Consequently, they argue, it is not possible to evaluate the accuracy of their glance detection algorithm.

The ability to determine another person's gaze direction has been a general area of interest. Studies into this topic suggest that the accuracy of such perception depends on where the observed person looks. For example, Bock et al. (2008) asked pairs of participants positioned at a 1m distance from each other to estimate where the other looked. One participant functioned as the 'sender' by looking at numbered tick marks located at a circle on a glass plate and the other participant functioned as the 'receiver'. On average, the responses were no more than 4° visual angle off from the target looked at by the sender. Note that in this situation the receiver looks directly at the sender and therefore these findings suggest that if drivers were to directly face the camera of an instrumented vehicle, subsequent annotation of glance direction will be highly accurate.

In naturalistic driving annotation, however, the face camera will often be mounted in the A-pillar or below the rearview mirror (otherwise their view would be blocked), resulting in an off-axis view of the driver. It is therefore important to establish how accurately humans can estimate glance direction under these conditions. Such off-axis glance perception was examined by Moors et al. (2015), who performed two experiments in which head and body orientation of the sender were manipulated. The first experiment focused on the frontal plane of a human character. When head and body were aligned, a 20° head rotation led to an underestimation of the gazed-at location. Additionally, an 'overshoot effect' was found: glance direction was systematically biased away from body orientation. In a naturalistic driving setting, these findings correspond to a perceived glance direction closer to the camera than in reality when the driver is looking straight ahead (i.e., head and body aligned). Assuming the overshoot effect, perceived glance will be overestimated if the driver looks marginally sideways (i.e., head and body misaligned).

The second experiment of Moors et al. (2015) focused on a sagittal view of a human character, in which the head orientation was manipulated in the downward and upward directions. A similar overshoot effect was found as in the first experiment. Objective looking angles were consistently overestimated, and the overestimation was larger for larger head orientations. In addition, an offset to the overshoot effect was found in that even a 0° looking angle (i.e., straight ahead) was judged as slightly downward. Projected again on a naturalistic driving setting, these findings suggest that a vertical overshoot in perceived glance direction may occur when drivers are looking up or down (e.g., a glance towards the speedometer may be misinterpreted as a glance towards the steering wheel). Vertical overshoot effects may depend on the driver's height: a tall driver may not need a large head tilt to look into the rearview mirror, whereas a short driver likely does. Moors et al.'s (2015) findings therefore suggest that manual annotation may show systematic biases, which depend on the head and eye gaze direction of drivers, and their height.

Another indication of possible issues with human annotation of glance behavior is found in the study by Ahlstrom et al. (2015), who used video recordings of real road driving, to investigate the association between self-reported subjective sleepiness (SRS) and post hoc observer-rated sleepiness (ORS). In a first experiment, novice observers were instructed to rate the level of sleepiness based on measures such as eyelid closure duration and blink frequency. A poor match between SRS and ORS was found, as well as a low inter-rater agreement. In a second experiment, pairs of video segments were presented: one featuring an alert driver and the other featuring a sleepy driver. Experienced observers were instructed to select the video segment with the sleepy driver. Despite moderate inter-rater agreement (which, as indicated, seems to be normally assumed to indicate good accuracy), the average percentage of correctly assessed video segments was only 35%. Moreover, in two out of four video pairs all raters (i.e., perfect agreement) identified the wrong video. Such findings are not only indicative of a low validity of observer-rated sleepiness, but also give rise to the question whether good agreement also means good accuracy, and how accurate human annotation is in general.

Research aim and paradigm

The aim of the present study is threefold: 1) assess the ability of annotators to accurately and reliably judge the direction that drivers are looking, 2) determine whether high agreement between annotators is always associated with high accuracy, and

3) explore whether clustering of glance locations into larger glance regions improves annotation accuracy. Such clustering may be based on what distinction is needed for a certain study (i.e., a top-down, or theory-driven approach), or be derived from the data (i.e., optimize accuracy for a given number of clusters; i.e., a bottom-up, or data-driven approach).

The first aim focuses on the annotation of glance and gaze behavior, but the results are likely to apply to glance behavior for activities other than naturalistic driving, including gaze behavior in infants (e.g., Farroni et al., 2004; Hood et al., 1998), where the use of eye trackers may interfere with naturalistic gaze behavior. The second and third aim relate to various types of human annotation, including the annotation of gaze direction in data from mobile eye trackers (e.g., Ioannidou et al., 2017), or the annotation of where participants take hold of objects in grasping movements (e.g., Hermens et al., 2014), where there may be a tendency to assume that agreement between annotators implies accurate annotation, and where detailed regions may be used that annotators cannot actually distinguish between. By using stimuli for which it is known where the actor is looking (by instruction), the accuracy and reliability of the human annotation can be directly tested. A similar paradigm may then be used in other contexts.

In service of these research aims, a set of stimuli with known glance directions (i.e., a true 'ground-truth') was constructed based on Naqvi et al. (2018), who instructed drivers to glance at a series of predefined locations inside a stationary truck cabin. To verify the predictions based on Moors et al. (2015) and to mimic the natural variability of driver height in naturalistic driving studies, glance behavior was collected for three drivers with different heights. Whereas most naturalistic driving studies involve one or two annotators per stimulus, we asked a larger group ($N = 10$) to obtain better estimates of the accuracy per location and the amount of agreement between annotators. To relate the data of ten annotators to the more common situation where only two annotators are used, a Monte Carlo approach was adopted, randomly selecting two annotators on each run of the simulation to examine the situation where only two annotators would have been involved.

Recent work has suggested that glances are guided by a need for information intake above visual saliency (Henderson and Hayes, 2017), and consequently, it can be assumed that information retrieval will be guided by the driving context and future actions. When annotating glances in naturalistic driving data the annotator may make use of this relationship. Imagine a scenario in which a cyclist is about to undertake a truck in an adjacent cycle lane, prior to making a right turn at an intersection. The right blind spot camera shows the presence of the cyclist, while at the same time, the face camera shows that the driver is looking to the right. An annotator may infer that the driver is looking at the right blind spot mirror to decide whether he or she should brake. In reality, however, the driver may have been looking through the right side of the front window, scanning for traffic ahead. Thus, an incorrect interpretation of the driver's glance direction may result from assumptions by the annotator on the driving context. In the present study, we aim to avoid such possible confounds of the driving context, by presenting annotators with just the view of the driver, to isolate glance estimation accuracy from interpretation of the driving context. A consequence of this approach, however, will be that scores on accuracy and inter-rater reliability for each glance location should be viewed as a preliminary estimation for annotation in actual naturalistic driving studies, and that future studies will need to address the role of context.

An important practical question is how often truck drivers check the blind spot mirrors, because this is a common cause of accidents between a large vehicle (a truck) and an often less protected cyclist, often leading to serious injuries or fatalities (Prati et al., 2018). Because of this practical importance, the present study focuses on truck drivers. Some indication of what to expect in terms of results can be derived from Fridman et al. (2016), who investigated glance behavior in passenger cars, showing annotators images of actors. This work shows that glances at the left/right window and at the left/right blind spot area are often too similar for accurate annotation, and consequently, the authors subsequently relied on generalized left and right glance regions for their analysis. Although for trucks a more fine-grained annotation than simply coding for leftward and rightward glances is often needed, because the different mirrors in trucks focus on different directions, the present study determines whether larger clusters can be used for annotation in trucks, where regions span larger visual areas both in terms of glance direction and visual angle. For example, when a driver looks left, it can be concluded that they did not check the blind spot mirror on the right. Therefore, the frequency of leftward glances will provide a lower estimate of how often the driver did not check the right blind spot mirror. The frequency of rightward glances similarly provides an upper estimate. Likewise, if one were to be interested in eyes-on-road and eyes-off-road, a binary distinction between those regions may also suffice.

Method

Participants

Ten participants (four males, six females, aged between 22 and 36) took part in the study, recruited by word of mouth and opportunity sampling. All participants were paid employees at the institute where the research was conducted (SWOV) and included one staff member, five interns, and four annotators serving on unrelated projects. None of them had extensive prior experience with annotation of glance behavior by truck drivers (the annotators just had started their role). All participants reported (corrected to) normal eyesight.

Stimuli

Videos were recorded inside a parked Volvo FH truck using a Sony RX100 for a view of the driver's face (sensor size: 25.41 mm, focal length: 18 mm, diagonal view: 52°) and a GoPro Hero3 for the cabin view (sensor size: 11.05 mm, focal length: 3.5 mm, diagonal view: 100°). Their position and field of view were chosen such that they mimicked the camera views provided by the face camera and cabin view camera in the UDRIVE truck database. For the same reasons, the resulting videos were converted to grayscale.

Twenty glance locations were defined, see Fig. 1: speedometer (location C1), center console (C2), driver seat (C3), passenger seat (C4), cabin roof (C5), cabin top left (C6), cabin top center (C7), cabin top right (C8), front window straight (F1), front window center (F2), front window right (F3), front blind spot mirror (F4), left window (L1), left side mirror (L2), left blind spot mirror (L3), right window (R1), right side mirror (R2), right blind spot mirror (R3), right pedestrian mirror (R4), and right door lower window (R5).[1]

Most glance locations correspond to those previously used in studies with passenger cars (e.g., Seppelt et al., 2017). Exceptions are the mandatory blind spot mirrors in trucks, the driver seat, and the passenger seat (C4). The latter two were included, because previous naturalistic driving studies on task distraction have shown drivers placing food, drinks, mobile phones, as well as travel documents in these locations (Carsten et al., 2017). Furthermore, glance location R5 corresponds with a hypothetical window in the lower part of the passenger door. Although this window is not present in the Volvo FH truck used for the videos, the Volvo FL trucks that were used in the UDRIVE study do include such a window. Finally, glance locations in the upper part of the cabin (C5 to C8) were selected because some trucks include equipment at these locations that draws visual attention (e.g., a display providing a view on the blind spot).

Three drivers were recruited to create the videos that served as the stimuli: a tall driver (standing height: 187 cm, sitting height: 96 cm), a medium tall driver (standing height: 176 cm, sitting height: 84 cm), and a short driver (standing height: 165 cm, sitting height 80 cm). Numbered sticky note were placed at the center of each glance location. The drivers were verbally instructed to briefly glance at a specified position before returning their gaze to the straight forward direction (i.e., glance location F1 in Fig. 1) and one of the other nineteen glance locations. The drivers were told to fixate their gaze at the Post-it for 2 s before returning to location F1. When all locations were looked at once, the sequence was repeated in a different order. The resulting videos were synchronized, placed side by side, and cut into shorter sections each showing one glance shift to a target location and back to the baseline position (see Fig. 2). In total, 114 such smaller videos were created showing two repeated glance shifts towards one of 19 locations for three drivers.

[1] One of the reviewers correctly suggested that measurements of the truck, in terms of the rotation angle required to look at these regions, would have been useful to report. At the time of the revision, however, it was not possible to obtain such measurements at a truck dealer (and no access to the original vehicle), due to the COVID-19 pandemic. Such measurement will therefore be absent.

Apparatus

Two 23-inch flat screens were used to present the videos and the images showing the regions. The OpenSesame software (Mathôt et al., 2012) was used to present the videos in a randomized order and allow participants to indicate when they were ready to give their response. To indicate the possible responses to the participant the image in Fig. 1 was used, showing each of the glance locations on the second screen. One image showed the same truck with lines indicating the location. The other image was a graphic illustration of the outside of the truck with the different labels.

Procedure

Participants were tested individually. They were seated in front of the screens and were asked to look at short video clips of a driver inside a truck looking at various areas in and outside the truck (Fig. 2), and simultaneously, images of the coding of the locations (Fig. 1). Video clips were repeated until the participant pressed a key on the keyboard to indicate to be ready to provide their response. Participants gave their response by speaking out the letter-digit combination of the area. These responses were immediately entered into a spreadsheet by the experimenter who sat next to the participant. The experiment took around 30 min to complete. Participants received no specific training to perform this task. This was done purposely to more closely mimic the annotation process in naturalistic driving research, where training with stimuli with known glance directions is not commonly used.

Data analysis

Responses and the actual glance locations were entered into a spreadsheet and loaded into R for further processing using the *dplyr* (Wickham et al., 2019) and *tidyr* (Wickham & Henry, 2019) packages for data wrangling and *ggplot2* (Wickham, 2016) for data visualization. Standard errors used for error bars in data plots were computed using the *boot* package (Canty et al., 2012; Davison & Hinkley, 1997). To compute Krippendorf's alpha (De Swert, 2012; Krippendorff, 2011), measuring agreement between participants, we used R's *irr* package (Gamer et al., 2019).

A mixed effects logistic regression model was used to examine the effect of driver height on accuracy, using the

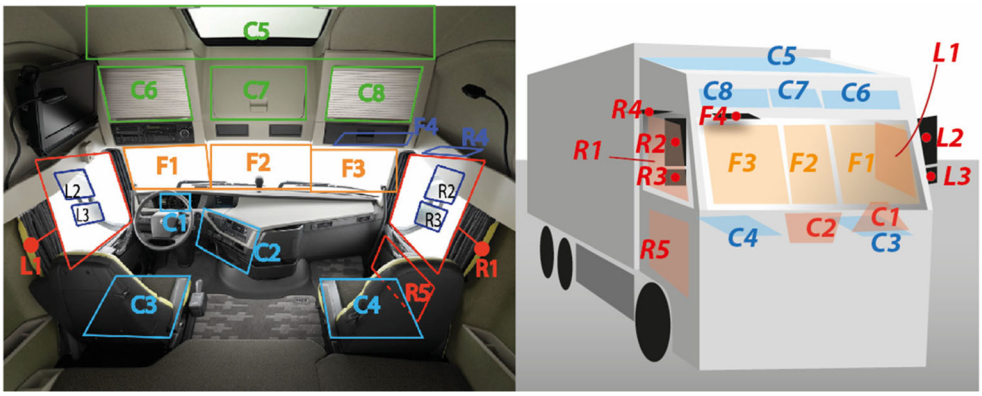

Fig. 1 Twenty glance locations from a cabin perspective (*left panel*) and an outside perspective (*right panel*). Prefixes: L = Left, R = Right, F = Front, C = Cabin

lme4 package (Bates et al., 2015). The choice for a mixed effects model, rather than a standard logistic regression, was made because there were two responses per glance location for each participant. A standard logistic regression would have assumed these two responses to be independent. The statistic for this test is a Chi-square value (results from a likelihood ratio test comparing a model with and without the effect of interest).

To examine possible clusters in participants' responses, the number of participants that gave a certain response for a certain target location was converted to a distance measure for that [correct response – given response] combination by subtracting the observed frequency from the maximum observed frequency (i.e., a large agreement means a smaller distance, see also Coxon & Jones, 1972). The resulting distance matrix was then made symmetric by averaging the distances at opposite locations from the diagonal. This distance matrix was then submitted to a hierarchical clustering analysis using the *hclust* method from base-R, using average linkage (Ward and complete linkage gave similar results, whereas single linkage led to the typical chains structure; Aldenderfer & Blashfield, 1984) and multi-dimensional scaling using the *cmdscale* method from base-R.

Distances were also analyzed using the R *network* package (Butts, 2008; Butts, 2015), which creates a visualization of the network structure between responses by plotting edges and nodes in such a way that clusters can be easily visually detected.

Results

Overall agreement and accuracy

As a first indication of the consistency of the responses of the participants, we computed Krippendorff's alpha, which provides a measure of congruency across more than two annotators. Across all drivers and glance locations a Krippendorff's α value of 0.39 was found, whereas per driver, values of 0.37 (short driver), 0.40 (medium height driver) and 0.40 (tall driver) were found, all well below the 0.67 threshold at which results are still considered acceptable (Krippendorff, 2004).

Another method to examine the extent to which participants' responses agree, is by examining the largest percentage of the same response per response category (not necessarily the correct response), shown in Figure 3a. This shows that for some locations the overlap of the most common response can be as high as 80% (annotators and repeated responses, e.g., driver seat, dashboard speedometer, left blind spot mirror), whereas for other locations the highest overlap drops to close to 25% (e.g., front window right, front window center). In

Fig. 2 Video still of the medium-tall driver looking at the right pedestrian mirror (glance location R4) after an instruction from the experimenter. Image published with permission of the participating driver

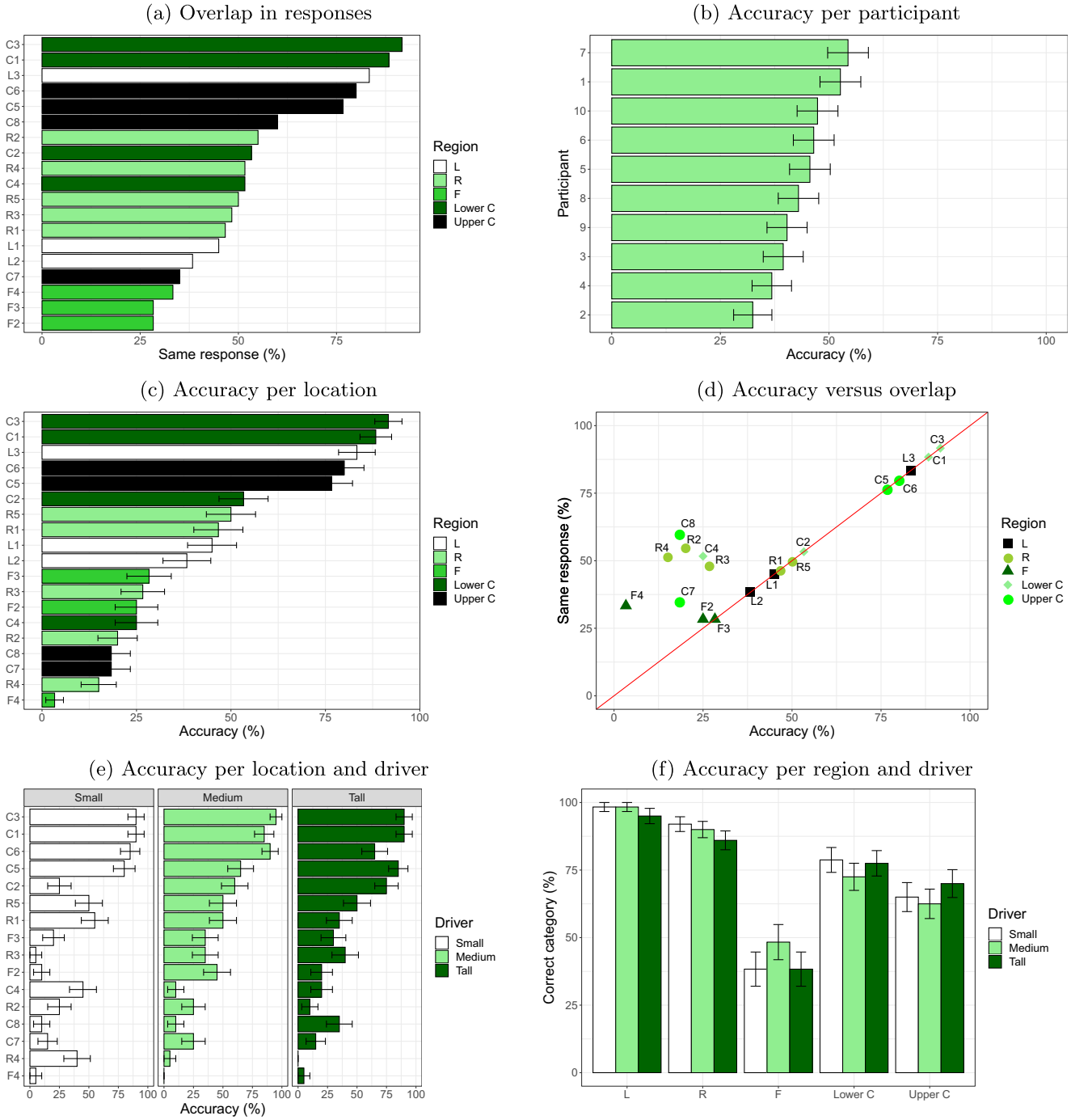

Fig. 3 **a** Highest percentage in overlap between responses for each glance location. **b** Accuracy per participant. **c** Accuracy per glance location. **d** Comparison of accuracy and overlap per glance location. **e** Accuracy per glance location and driver. **f** Accuracy per glance region and driver. *Error bars* show the standard error of the mean obtained with bootstrapping. Prefixes: L = Left, R = Right, F = Front, C = Cabin

part, this overlap is linked to accuracy: If many participants are able to provide the correct response, the overlap can be expected to be high. Participants, however, may also systematically provide an alternative response, meaning a high overlap and low accuracy.

To examine the extent of agreement between overlap and accuracy, Fig. 3b–d plots the accuracy per participant, accuracy per location, and the accuracy per location against the overlap per location. Generally, participants were not very accurate (Fig. 3b; best participant with an accuracy slightly above 50%), but there were no participants that clearly stood out. This agrees with feedback from participants, who indicated not to be very confident about their responses. Locations also differed in how accurately they could be reported (Fig. 3c), with

accuracy ranging from almost zero (front blind spot mirror) to close to 90% (driver seat, dashboard speedometer). Several cabin locations could be well identified, but this was not the case for all cabin locations, as some were among the poorest scoring locations (e.g., cabin top right, cabin top center, passenger seat).

Accuracy versus agreement

Figure 3d indicates that many glance locations showed a clear correspondence between overlap and accuracy (e.g., locations C5, C6, L3, C1, and C3), where participants agreed simply because they each chose the same correct response. For some locations, however, overlap was larger than accuracy, meaning that participants tended to consistently choose a different location than the correct one (glance locations C4, C7, C8, R2, R3, R4, and F4 above the diagonal in Fig. 3d). As annotation studies typically take agreement between annotators as an indication of a reliable annotation, this observation is particularly important: High agreement does not necessarily mean accurate annotation, and it is difficult to predict when high agreement means accurate annotation and when it does not.

The analysis so far has considered all ten participants. In naturalistic driving studies, however, it is common to only have two annotators code each stimulus, as opposed to ten. To relate the findings of Fig. 3d to common practice, we performed a simulation in which repeated random draws of two annotators were made, for which we calculated agreement and accuracy. This Monte Carlo simulation used 100 repetitions per stimulus (more draws were significantly slower to perform while not substantially changing the results), corresponding with 200 draws per glance location (because there were two videos for each location). The simulation showed that, the chance of perfect agreement is 42.8%. For all draws where both participants agreed, the chance of a correct response is 63.4%, meaning that agreement meant an incorrect response in the remaining 36.6% of the cases.

Figure 4 shows that for only a small subset of locations (C1, C3, C5, C6, and L3) the chance of agreement (i.e., 'both correct' plus 'both incorrect') is larger than 50%, whereas the chance of disagreement is larger than 50% for the remaining 14 glance locations. Practically, this means that in many cases, a third annotator would be needed to resolve the disagreement (or more, if this third annotator disagrees with the first two annotators). Moreover, in case of agreement, incorrect judgments are more likely than correct judgments for a large number of glance locations (C4, C7, C8, R2, R3, R4, F2, F3, and F4). Note that these findings are generally in line with the results presented in Fig. 3d (i.e., glance locations above the diagonal).

Effect of driver height on accuracy

Videos of three different drivers were used, with different heights. Figure 3e shows that there was some variation in how accurately participants could report the glance direction

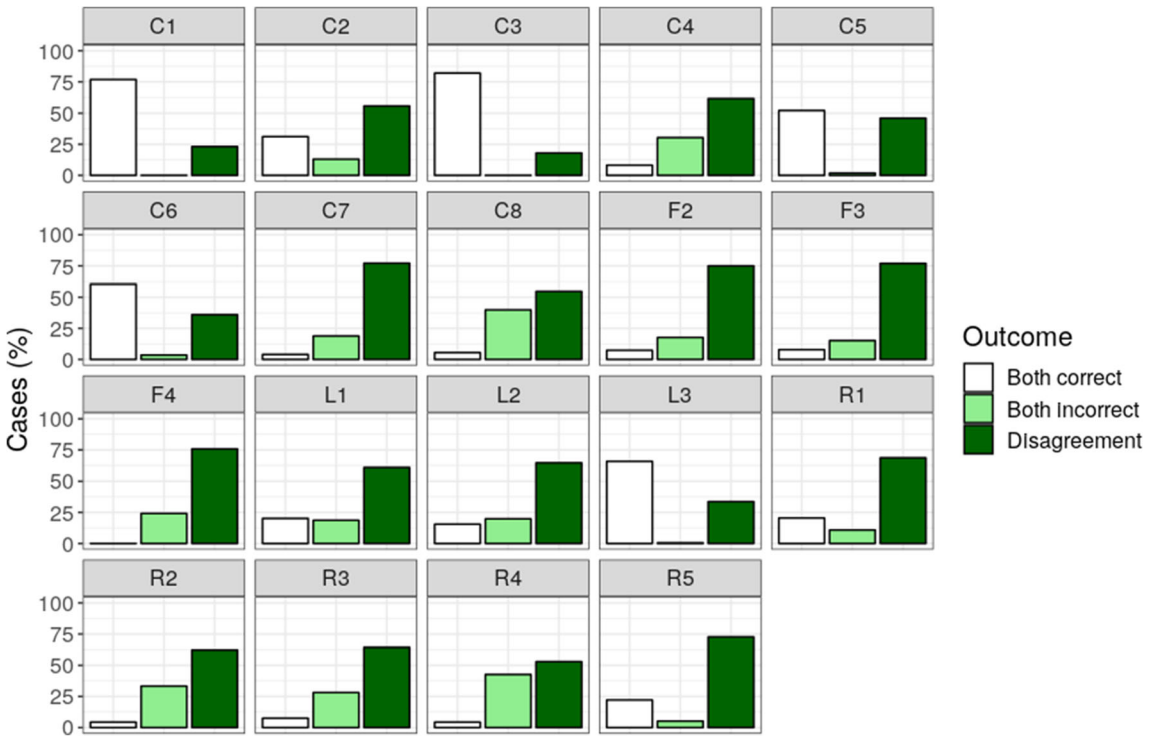

Fig. 4 The average chance that two randomly selected annotators agree and whether they were correct, as a function of glance location

of these drivers for each location. A mixed effects logistic regression found a significant effect of driver on accuracy (after Bonferroni correction for multiple comparisons) for the right pedestrian mirror ($\chi^2(2) = 15.86$, $p < 0.001$) and the center console ($\chi^2(2) = 24.48$, $p < 0.001$). Test results for the remaining comparisons are shown in Table 1 in the appendix, suggesting that the effect of driver height is driven by the difference between the accuracy for the short and tall drivers. For the right pedestrian mirror, taller drivers led to a lower accuracy, whereas for the console, the direction of the effect is reversed. Note that the effect of driver is sometimes different for adjacent regions (e.g., R4 and F4). Given the relatively small numbers of observations (60 overall, when comparing the three drivers), very strong conclusions of the effects of driver height should not be drawn on the basis of these data (it was not the focus of this study, but given the results, further investigation is warranted).

Effect of region clustering on accuracy

The results show that generally participants were not very accurate, with some glance locations having very low annotation accuracy. A possible reason may be the large number of response categories: People may not be able to distinguish glance direction with this level of detail, and when uncertain, the chance of guessing correctly is lower with a higher number of categories. It may therefore be beneficial to cluster glance locations into larger glance regions that annotators can distinguish better, like past studies of glance direction in naturalistic driving have done. Two approaches may be used in this context: (1) Clustering on the basis of a priori categories, and (2) a posteriori clustering, for which we use clusters observed on the basis of confusion counts between categories.

A-priori clustering

A first approach is to use the five main regions: Front ('F' prefix), the upper part of the Cabin ('C-upper' prefix), the lower part of the cabin ('C-lower'), Right ('R' prefix), and Left ('L' prefix). An overall accuracy of 75% is found for these five regions (see Fig. 3f for the per-region accuracies), which is better than chance (20%), but likely to be insufficient for application in naturalistic driving studies or to train automatic detection algorithms.

A posteriori clustering

Even with the five large clusters, reliable accuracy was not achieved for all regions. The a priori grouping may therefore not be optimal: probably participants confused glance directions at the edges of these large regions. To determine whether it is possible to obtain better clusters (higher overall accuracy) on the basis of the data, we made use of unsupervised learning techniques to a posteriori cluster the various glance locations into larger clusters. The exact procedure when applying these techniques is described in the appendix. Input for these techniques are the confusion counts, which indicate how often one region was 'confused' with another region.

These counts are shown in Fig. 5, which plots the frequency of responses for each combination of a correct and a given response (with brighter colors indicating more frequent combinations). The bright areas on the diagonal indicate correct responses, but there are also various brighter off-diagonal areas, which indicate areas that are often confused. The most frequent confusions are R4 responses for C8 glances (60%), R2 responses for R1 glances (55%), C4 responses for R5 glances (52%) and R4 responses for R1 glances (also 52%). These frequencies also imply that some responses occurred often (e.g., R1: 166 times, L3: 90 times), and some responses were avoided (e.g., C7: 21 times, and F2: 26 times). Confusions were also not always symmetric. For example, glance location C8 is often responded to with R4, but R4 was not often responded to with C8. Likewise, glance location F2 is sometimes responded to with the right-side mirrors R1, R2, and R3. The opposite effect, where the right-side mirrors are responded to with glance location F2, was not found. We inspected the videos to examine whether there were differences in glances to F2 and those to R1, R2, and R3 with respect to the involvement of head movements that could explain the asymmetric nature of the errors for these regions. Glance shifts to these four regions all involved a combination of a gaze-shift and a head turn, but the head turn was larger for the R-regions than for the F2 region. The asymmetric confusions between F2 and the R-regions therefore seem to be indicative of an overshoot effect, rather than being due to a difference between eye-gaze and head-turn glances.

Three different a posteriori clustering solutions were found, depending on the method used and how the results are interpreted (see appendix for details). A clustering of the

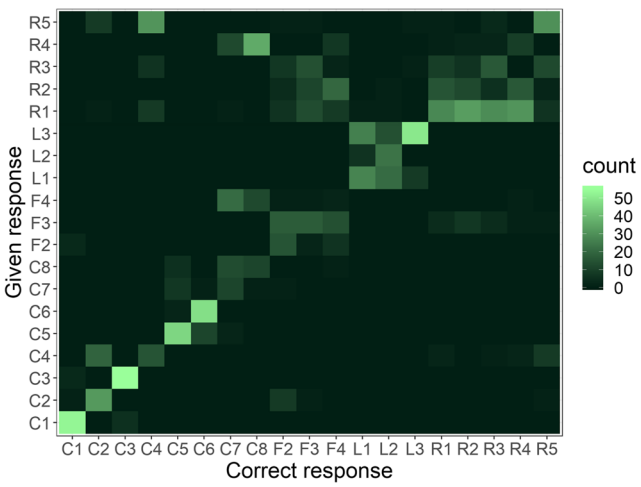

Fig. 5 Confusion counts between the various glance locations

various glance locations into 11 small clusters ({L3}, {L1, L2}, {C1, C3}, {C4}, {R4}, {R1, R2}, {R3, R5}, {F3, F4}, {C2, F2}, {C8}, {C5, C6}) gives an overall accuracy of 60% (cf., with the original labels: 44%). When a broader five cluster solution is used ({C1, C3}, {C2, F2, F3, F4}, {C5, C6, C7, C8}, {L1, L2, L3}, {R1, R2, R3, R4, R5, C4}) an accuracy of 80% is found, which is higher than the 75% found for the a priori clustering. In terms of spatial proximity (Fig. 1) the five clusters in this solution also make sense. A third a posteriori clustering (see 'MDS' in the appendix) has three clusters (L-regions, R-regions, other regions). For this clustering an overall accuracy of 82% is obtained, only marginally better than with the larger clusters of the hierarchical clustering analysis (five rather than three clusters) – but set against a substantially higher chance level when annotators guess (1 in 3, rather than 1 in 5).

Discussion

Naturalistic driving studies often make use of manual annotation to determine the momentary glance direction of drivers from video images. When using such annotation, it is often assumed that when annotators agree, the annotation will be correct. In this study, we tested this assumption by showing videos of truck drivers looking at predefined regions inside and outside the truck and asking participants to name these regions. Our five main findings are: 1) there are large differences both in terms of accuracy and percentage agreement across glance locations, 2) some glance locations show a significant effect of driver height on accuracy, 3) average accuracy was consistently low across the annotators, 4) some glance locations showed a very low accuracy, despite a moderate degree of agreement, and 5) clustering glance locations into larger glance regions improves accuracy. Next, we discuss the implications of our findings, followed by limitations of the study and recommendations for future research.

Implications

Implications for automated glance extraction

A major implication concerns the automated extraction of glance directions. To develop a system to automatically annotate glance direction from video material, a training set is needed with video images coded for the actual glance direction (for supervised learning). Several studies either explicitly or implicitly use manual annotation as the 'ground truth' for the development of image processing algorithms (e.g., Fridman et al., 2016; Tawari and Trivedi, 2014; Vora et al., 2017). The general assumption for such manual annotation appears to be that if two or more annotators are in agreement, the classification must be correct. Our findings demonstrate that such an assumption cannot be automatically made. On average, 37% of judgments with perfect agreement between two annotators were incorrect. Sometimes a third annotator is used to resolve the disagreement, but it is not unlikely that this third annotator would disagree with the first two annotators, and even if the third annotator would agree with one of the first two annotators, the 'majority vote' is not automatically the correct judgment. Therefore, studies on image processing algorithms that fail to collect a true ground-truth may show a high internal validity (with regard to perceived glance behavior), but not necessarily a good external validity (with regard to actual glance behavior).

Implications for studies on traffic safety

While our study was conducted on glance behavior collected in the absence of a driving context (i.e., lacking views of the exterior of the vehicle, and recordings made inside a static truck), the present results can be interpreted as a first indication that human annotation of glance behavior may not be as accurate as often implicitly assumed, and that inter-coder agreement does not automatically mean accurate coding. The impact of our findings will also depend on the aim of the study (i.e., which glance location or region needs to be coded with high accuracy). For example, in studies on speeding it may be relevant to confirm whether the driver has visually inspected the speedometer. With an average accuracy of 88% across ten annotators, it is likely that a true glance towards the speedometer (C1) is perceived as such, even in the absence of visual cues such as optic flow visible on the forward-facing cameras indicating the speed of the vehicle.

In contrast, studies on driver distraction often require being able to distinguish between off-road and on-road glances (e.g., Lin et al., 2019). For off-road glances, high accuracy was found for the driver seat (C3) and speedometer (C1), but not for the passenger seat (C4, related to driver distraction), which was often confused with the passenger door window (R5). This confusion between glances at the passenger seat and passenger door could be explained by the fact that drivers looking at the passenger door window may need to look just across the passenger seat. Note that this confusion becomes irrelevant if the truck at hand does not feature a passenger door window. Furthermore, a high annotation accuracy was found for the center console (C2), but only with the tall driver. This effect can be explained by the fact that a taller driver requires a larger downward facing head orientation, which is more easily detected than a smaller amplitude downward glance towards the same location by a short driver.

If larger clusters can be used, higher accuracy can be achieved. For example, when all cabin glance locations are clustered into two categories (upper and lower), and the remainder of the regions are kept, an overall accuracy of 75% is obtained. This accuracy may still be below what is required for a naturalistic driving study or the use of the annotated data

for the development of automatic detection algorithms. A possible reason may be the confusions with glance locations that are not part of the cabin area, such as between the cabin top center (C7) and the cabin top right (C8) on the one hand, and the front blind spot mirror (F4) and the right pedestrian mirror (R4) on the other hand. Taken together, one must conclude that the accuracy of off-road glance annotation will depend on (1) the glance locations that are selected as part of the study, (2) whether glance locations are clustered, and (3) possibly the driver's height. For on-road glances, the accuracy for locations F2 (front window center) and F3 (front window right) was below 25%, as these locations were often confused with the right-side mirrors. Even when glance locations F2 and F3 are clustered, the resulting accuracy will still be below 50%. The confusion of these locations was unidirectional: the right-side mirrors were never confused with F2, and only occasionally with F3. This is indicative of an overshoot effect as described by Moors et al. (2015). The data, however, also show that left and right glance directions can be distinguished with high accuracy (confusions between left and right only account for 1.12% of all left and right region responses). So, for studies that only distinguish between left and right glance directions (e.g., infant studies of gaze following), the present findings should not be a cause of concern.

The findings also have important implications for studies on blind spot accidents. The results show that it appears to be impossible to accurately distinguish between separate mirrors on the passenger's side. Although the accuracy for the right pedestrian mirror (R4) was significantly larger with the short driver, likely due to a larger vertical head orientation, the accuracy did not surpass 40%. In fact, across all drivers, the accuracy was lower than 50% for all glance locations on the passenger side. Annotation with a broader category, however, yielded better results. Both an a priori categorization and a categorization based on multidimensional scaling showed an accuracy of over 80% when all glance locations on the passenger side are grouped, possibly removing the effects of confusions between glances to the different mirrors and glances through the right window (R1). Studies on blind spot accidents should therefore focus on the presence or absence of sideways glances on the passenger side. When a driver looks sideways, they potentially checked one of the mirrors, but one should refrain from conclusions about which mirror was checked, and also take into account that the number of sideways looks reflects the maximum number of times the driver checked the blind spot mirrors on this side. If the driver does not look sideways, then one can conclude the blind spot mirrors on this side have not been checked. Note that the front blind spot mirror (F4) showed the lowest accuracy of all glance locations. Therefore, aforementioned remedy may help for studying behavior related to accidents where other road users are hit at the side of the truck, but not those where the impact occurs at the front of the truck.

Limitations

Controlled setting

Videos that served as stimuli were recorded inside a parked truck. Such a setting, however, may have led to different glance behavior compared to naturalistic driving. A possible alternative would have been to use a think-aloud protocol in a naturalistic driving situation, where the driver at each instant indicates where (s)he is looking. Think-aloud protocols have been used in real traffic research, for example to understand trust ratings in automated vehicle technology (Ekman et al., 2019), and to measure situation awareness (Key et al., 2016). In both cases, verbal utterances were based on (prospective) interactions with other road users, such as passing a pedestrian crossing or taking a roundabout. There are several reasons why we did not opt for such a strategy. First, the use of a think-aloud protocol for glance direction may lead to non-natural glance behavior. People tend to move their eyes about three times per second (Rayner, 1998). To avoid having to name each of these glances, drivers may try and reduce their glance shifting frequency. There are some indications adjustments in glance behavior may take place. In Kircher and Ahlstrom (2018) drivers used more time on mirror glances when driving with a concurrent think-aloud protocol, compared to a baseline driving condition where no protocol was used. The authors argue that mirror glances, which are typically short, increase in duration by mentioning them and/or by becoming aware of automated glance behavior. Second, and related, self-reported glances collected through a think-aloud protocol are likely incomplete. With a frequency of around three glances per second, it is almost impossible to keep up verbally. Third, asking drivers to drive normally and memorize their glance directions is also unlikely to work, as this would require a large number of directions to be stored, beyond the capacity of normal human working memory (Ericsson & Simon, 1980). Some of the locations will not be glanced at often, and therefore quite a large interval of driving and glance behavior may need to be collected before all regions are covered. In contrast, asking drivers to look at a series of Post-its unequivocally yields the desired collection of glance directions. Finally, related to previous points, think-aloud protocols may also impact driving behavior and possibly performance (and therefore affect safety). Given the rate with which drivers typically change their glance direction, a think-aloud protocol would invite drivers to talk continuously. In the meta-analysis of Caird et al. (2018) talking with a passenger was shown to result in slower reaction times, decreased hazard perception, and more collisions, compared to baseline conditions without talking. These negative effects on driving performance may be mitigated partially by telling drivers to pause verbalization during challenging traffic situations. This, however, will restrict collected glance behavior to less than challenging driving conditions. Consequently, the collected glance behavior may not be representative of the often-

inspected safety critical events in typical naturalistic driving studies.

A second limitation is that we only provided participants with video images of the driver, stripping the analysis of further context. In naturalistic driving studies, annotators usually have several outwards facing cameras at their disposal, which can be used to infer what drivers may be looking at, and consequently, to select a glance location. It is unclear to what extent annotators actually use these images, which could be addressed in further research using eye tracking to determine which images and where in the images annotators look when coding for glance behavior in drivers. Such a study would possibly inform any differences between coding by different annotators, and what cues annotators use to complete their task, which may also benefit training purposes.

A further limitation may be that, in contrast to naturalistic driving, the fixation interval of 2 s in our study may have been rather long (with the knowledge that human shift gaze about three times per second). For annotation, a relatively long fixation duration can be expected to improve annotation accuracy. Given that annotation accuracy was rather low, we therefore expect that for actual naturalistic driving data, annotation accuracy may be even lower.

Another difference compared to naturalistic driving is that drivers may not always glance back to the front direction (i.e., glance location F1), as was the case in the videos used. As for the longer duration, changing this feature to more realistic gaze behavior can be expected to reduce annotation accuracy due to stronger confusion between glanced at regions.

Furthermore, the truck was situated in a well-lit garage, which does not represent situations of high contrast due to sunlight or extreme darkness when driving at night. Therefore, it is plausible that the accuracy of annotation in real driving conditions will be lower than then values reported here.

A related limitation is that the present study focuses on the direction of gaze (foveal vision), whereas it is known that in naturalistic driving extrafoveal vision plays an important role (e.g., Wolfe et al., 2017). Note that this limitation is not unique to studying glance behavior using human annotation. The limitation also plays a role when glance directions would have been measured using eye trackers, which do not measure extrafoveal, covert attention either. Extrafoveal vision differs from foveal vision in various aspects (for an overview, see Rosenholtz, 2016), including lower visual acuity (Rayner, 1978), the phenomenon of crowding (difficulties differentiating features of objects with neighboring objects present; Whitney & Levi, 2011), and altered color perception (Sivak et al., 2000). Moreover, it can be demonstrated that extrafoveal processing is used in day-to-day tasks such as reading. For example, a moving window that is dynamically adjusted depending on where the reader fixates, can reduce reading speed, depending on the size of the window and the features of the text outside the moving window (e.g., whether the text still contains spaces, or consists of letters that are similar; Rayner & Pollatsek, 1994). There are also indications that foveal vision is specifically used in situations where shape information is difficult to extract from extrafoveal vision. For example, when asked to report the direction of a pair of eyes (embedded inside a face, likely to cause visual crowding) or a pointing hand presented away from fixation, participants more often make an eye movement to the pair of eyes than the hand when allowed to, and perform more poorly in a direction discrimination task of the pair of eyes than the pointing hand when not to allowed to look at the stimuli (Hermens, Bindemann & Burton, 2017). Peripheral vision, however, needs to serve a purpose. In the indoors setup used here, peripheral vision may be of less importance, because of the absence of a surround of interest to visually inspect. This may have affected glance behavior in the drivers.

In the present work, we focused on trucks, because of the importance of glance behavior in blind spot accidents. The results cannot fully be extended towards driving in passenger cars. Cars have fewer areas of interest (e.g., no blind spot mirrors) in the same glance direction and therefore, annotators might have fewer difficulties distinguishing what car drivers look at. However, even for passenger cars, annotation often involves a task that is more complex than determining whether a driver looks right or left. For example, it can be of interest to know whether the driver looks through the window, at passengers or at the right mirror.

The overshoot effect observed in the present study raises the question to what extent forward right glances have been misinterpreted as right window glances in studies with passenger cars. For example, the study of Belyusar et al. (2016) on the effect of digital billboards on glance behavior may have missed forward right glances (i.e., underestimating the actual number of glances in this direction). In the study by Seppelt et al. (2017) on using glance metrics to distinguish between crashes and near-crashes, on-road glances (i.e., including forward right glances) may have been mislabeled as off-road glances due to the overshoot effect.

Baseline glance direction

Actors were asked to always look back to the center (F1) position after glancing at each of the instructed positions. The reason for using F1 as the reference position, was that drivers tend to glance at this center position for a substantial amount of time during naturalistic driving (Fridman et al., 2016). In our study, we decided to show the full video clip, showing the glance shift from F1 to the target position, and the glance shift back to F1. It is yet unclear whether this sequence best resembles actual annotation of naturalistic driving data, where annotators may slow down, reverse, or pause the play-back.

The latter situation, where the video is paused, resulting in the annotator watching a static image, may be more in line

with studies examining social attention. In these studies, participants typically watch either a static image of a person looking ahead, followed by a blank, followed by a static image of a person looking left or right (static stimulus condition; e.g., Bayliss, Pellegrino & Tipper, 2005; Hermens & Walker, 2010) or a video of a person first looking ahead, and then shifting their glance leftwards or rightwards (dynamic stimulus condition; e.g., Hermens & Walker, 2012; Rutherford & Krysko, 2008; Swettenham et al., 2003). It has been suggested that attention shifts following dynamic stimuli may be stronger, because the same neurons in the human brain may respond to both social (Perrett et al., 1985; Perrett et al., 1992) and biological motion cues (Oram & Perrett, 1994). A direct comparison between static and dynamic cues, however, did not show such a stronger attention shift for dynamic cues (Hietanen & Leppanen, 2003).

Studies on social attention have typically avoided asking actors to look back at the starting position to avoid the inhibition of return effect, where the cued direction is inhibited after drawing attention back to the center (after a certain delay; Posner & Cohen, 1984; Klein, 2000). These studies therefore provide less information about the consequences of asking actors to look back to center, also because they focus on automatic attention shifts, and less on estimating the direction of a perceived glance shift. Social attention studies have also largely focused on eye-glance shifts, with fewer studies examining whole-head glance shifts, whereas whole-head glance shifts tend to result in stronger attention shifts in the observer, particularly when the stimuli are looked at from peripheral vision (Burton et al., 2009; Hermens et al., 2017). Future studies should therefore examine what the effects of using dynamic versus static stimuli are, and whether showing the section of the video in which the actor looks back at the center position affects the findings.

Recommendations

This study shows that it is not advised to have blind faith in the judgment of annotators when annotating glance behavior, even when two annotators agree on the coding. Instead, the findings suggest annotation may be improved by providing annotators with a reference set of images or videos showing drivers' glance behavior that show drivers looking at predefined locations in the vehicle. Such a reference set of images or videos can aid the training of annotators (e.g., Chapman et al., 2008; Cabrall et al., 2018), and help annotators improve the accuracy of their annotations. Once high accuracy after such training is achieved, the resulting data can be used to develop video algorithms that automatically code what drivers look at (Naqvi et al., 2018).

All-in-all, our findings suggest that people perform poorly when annotating glance locations at a high spatial resolution, for many locations inside a vehicle. No participants clearly stood out in their performance, as is sometimes found in face recognition, where there are super-recognizers (Bobak et al., 2017; Robertson et al., 2016; Russell et al., 2009) or people particularly poor at recognizing faces (Damasio et al., 1982; McNeil & Warrington, 1993). Thus, it is unlikely that selecting another sample of untrained participants would have improved annotation accuracy. Improved annotation accuracy may be achieved after training (e.g., Chapman et al., 2008; Cabrall et al., 2018). A follow-up study on the effect of training is therefore warranted.

Finally, future studies could further address the generalizability of our findings towards annotation in naturalistic driving studies by showing annotators the driving context (e.g., views from other cameras, sound recordings), and by using glance behavior in actual traffic. The main challenge for constructing such stimuli will be to generate a data-set of known glance locations without interfering with either driving or glance behavior. Possibly the use of a single, high-end remote eye tracker may aid this purpose.

Conclusions

In naturalistic driving studies agreement between pairs of annotators is often seen as an indication of accuracy. The present study demonstrates that such an assumption is not automatically correct: annotators can be in agreement, but both use the incorrect label. Clustering of regions improves accuracy, but not to a level that would be considered sufficient for most naturalistic driving applications, particularly for regions further away from the camera. Correct annotation is important for training of algorithms with the aim of automatically detecting gaze direction, and it is therefore important to first experimentally test the annotation accuracy, as was done in the present study, before using such annotation for algorithm development.

Acknowledgements We thank Ron Schindler and Giulio Bianchi Piccinini for recording and sharing the videos of glance behavior of three drivers, and the drivers for their cooperation. This study was funded by the Dutch ministry of infrastructure. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Open practice statement The datasets and analysis scripts from the study are available in the OpenScience Framework repository, https://osf.io/5upfh/?view_only=2b7f284db18c4ad0b5159b679c8b38db. The study was not preregistered.

Author contributions Reinier Jansen: Conceptualization, Methodology, Supervision, Writing – Original draft, Writing – Review & Editing. Sander van der Kint: Software, Investigation, Writing – Original draft, Visualization. Frouke Hermens: Methodology, Software, Formal analysis, Writing – Original draft, Writing – Review & Editing, Visualization.

Appendix

Table 1 Statistics for comparisons of response accuracy per driver. Note: When a p-value equal to one is shown, this means that the number of correct and incorrect responses were the same across drivers

All three sizes			Short vs. Medium			Short vs. Large			Medium vs Large		
Location	Chi-square	P-value	Location	Chi-square	P-value	Location	Chi-square	*P* value	Location	Chi-square	P-value
C1	0.313	0.855	C1	0.23	0.632	C1	0	1	C1	0.23	0.632
C2	24.479	< 0.001	C2	10.78	0.001	C2	22.967	< 0.001	C2	3.058	0.08
C3	0.523	0.77	C3	0.537	0.464	C3	0	1	C3	0.537	0.464
C4	8.664	0.013	C4	7.001	0.008	C4	4.109	0.043	C4	0.876	0.349
C5	3.749	0.153	C5	1.661	0.197	C5	0.579	0.447	C5	2.84	0.092
C6	5.248	0.072	C6	0.286	0.593	C6	3.043	0.081	C6	4.857	0.028
C7	0.859	0.651	C7	0.63	0.427	C7	0	1	C7	0.63	0.427
C8	5.314	0.07	C8	0	1	C8	3.752	0.053	C8	4.188	0.041
F1	8.907	0.012	F1	9.026	0.003	F1	0.797	0.372	F1	3.805	0.051
F3	1.773	0.412	F3	2.128	0.145	F3	0.681	0.409	F3	0.158	0.691
F4	1.656	0.437	F4	1.412	0.235	F4	0	1	F4	1.412	0.235
L1	0.553	0.758	L1	0.175	0.676	L1	0.499	0.48	L1	0.119	0.73
L2	7.343	0.025	L2	6.904	0.009	L2	1.661	0.197	L2	2.63	0.105
L3	1.965	0.374	L3	0.233	0.629	L3	0.673	0.412	L3	2.196	0.138
R1	2.128	0.345	R1	0.103	0.748	R1	1.86	0.173	R1	1.616	0.204
R2	2.144	0.342	R2	0	1	R2	1.644	0.2	R2	1.644	0.2
R3	11.259	0.004	R3	8.009	0.005	R3	8.026	0.005	R3	0.161	0.688
R4	15.864	< 0.001	R4	7.792	0.005	R4	13.268	< 0.001	R4	1.412	0.235
R5	0	1	R5	0	1	R5	0	1	R5	0	1

A posteriori clustering of glance locations

Unsupervised clustering techniques were used to establish possible clustering of glance locations on the basis of the data. The techniques that we use work on a matrix of numbers that indicate the distance between each possible combination of locations. The task is therefore to create a measure that expresses this distance. Common methods for measuring or deriving distance measures are to ask participants to directly rate the similarity between stimuli, to score stimuli on a range of features and compute a measure such as the correlation, or to examine how often two stimuli are confused (Coxon and Jones, 1972). This first approach would require asking participants to compare videos for similarity, which will be hugely time-consuming. The second approach would require establishing features that describe features of each video. It would be unclear what features to define. We therefore adopt the third approach and use the number of times each response is confused with another.

Large confusion counts indicate that glance shifts looked similar to our participants. As clustering methods use a measure of dissimilarity rather than similarity, the confusion counts were inverted by taking the maximum pairwise count and subtracting the count for each combination of a correct and a given response. Clustering methods also assume that similarities are equal for A-B and B-A pairs. In a second step, the resulting distances between pairs (e.g., R5-C4 and C4-R5 distance) were therefore averaged in order to obtain a symmetric distance matrix assumed by the different techniques (cluster analysis and multi-dimensional scaling). Because establishing the dissimilarity matrix using the observed confusion counts requires several steps, it will be important to examine (1) whether the solutions from the various clustering methods make theoretical sense, and (2) whether different methods of clustering yield similar results (Coxon and Jones, 1972).

The first method that we used is hierarchical clustering, which is a bottom-up (agglomerative) approach that starts with the leaves, and successively merges leaves and clusters into larger clusters until the top of the tree is reached (Fig. 6a). The choice of which leaves or clusters to merge depends (in addition to the distance) on the linkage method, which indicates which distance to use between already formed clusters or clusters and leaves (e.g., the shortest, the longest, or the average distance). We here show the results of average linking, because other methods yielded a similar clustering (except for single linkage, which showed a long chain typically found for this method). The results show that left regions (L1-L3) and right regions (R1-R5) were grouped into different clusters, which makes sense from a theoretical perspective. The clustering solution also makes sense in terms of clustering cabin locations and frontal locations (Fig. 1).

Using the resulting tree, a posteriori clustering can be established. The tree can be converted to clusters by cutting the tree at a given height. Depending on where the tree is cut, different numbers of clusters are found. When the tree is cut at a low point, we find 11 small clusters ({L3}, {L1, L2}, {C1, C3}, {C4}, {R4}, {R1, R2}, {R3, R5}, {F3, F4}, {C2, F2}, {C8}, {C5, C6}). Cutting the tree at a higher point leaves five clusters ({C1,C3}, {C2,F2,F3,F4}, {C5,C6,C7,C8}, {L1,L2,L3}, {R1,R2,R3,R4,R5,C4}).

Hierarchical clustering trees do not provide a clear insight into how near the elements within each cluster are. Figure 6b therefore shows the results of a second method, which aims to establish the underlying dimensions that best represent the observed distances (multi-dimensional scaling, MDS). The solution is plotted using the two main dimensions, allowing for visualization of the result. The MDS solution suggests three main clusters: one on the top left of the map {L1, L2, L3}, one on the right {R1, R2, R3, R4} and one in the bottom left (the remaining locations). The left and right clusters agree with those found for hierarchical clustering, but MDS provides less information about how the Cabin and Front regions should be clustered.

Finally, Fig. 6c shows the results of a third method that uses graphs, which can deal with asymmetric distance matrices. It shows the same clustering of the L-areas {L1, L2, L3}, but no clustering of the R-areas. Instead, a cluster of C-areas can be seen on the left of the graph. The graph method therefore suggests a solution in between that of hierarchical clustering and MDS.

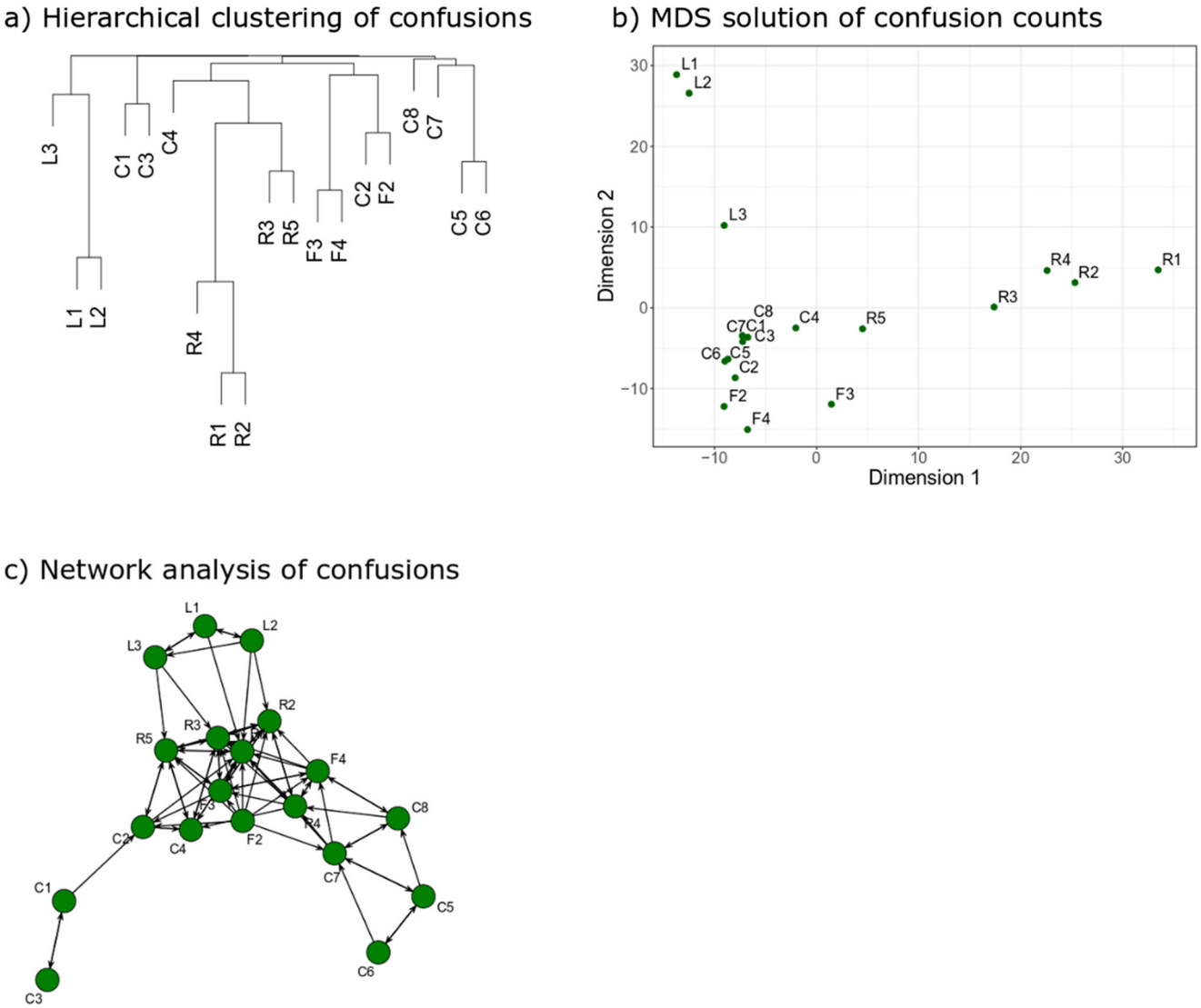

Fig. 6 **a** Hierarchical clustering solution using the confusion counts. **b** MDS result using the confusion counts. **c** Network analysis of the confusion counts

References

Ahlstrom, C., Fors, C., Anund, A. et al. Video-based observer-rated sleepiness versus self-reported subjective sleepiness in real road driving. *European Transport Research Review,*7, 38 (2015). https://doi.org/10.1007/s12544-015-0188-y.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills: Sage Publications

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bayliss, A. P., Pellegrino, G. D., & Tipper, S. P. (2005). Sex differences in eye gaze and symbolic cueing of attention. *The Quarterly Journal of Experimental Psychology, 58*(4), 631–650.

Belyusar, D., Reimer, B., Mehler, B., & Coughlin, J. F. (2016). A field study on the effects of digital billboards on glance behavior during highway driving. *Accident Analysis & Prevention, 88*, 88–96. https://doi.org/10.1016/j.aap.2015.12.014

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *The Quarterly Journal of Experimental Psychology, 70*(2), 201–217.

Bock, S. W., Dicke, P., & Thier, P. (2008). How precise is gaze following in humans? *Vision Research, 48*(7), 946–957. https://doi.org/10.1016/j.visres.2008.01.011

Burton, A. M., Bindemann, M., Langton, S. R., Schweinberger, S. R., & Jenkins, R. (2009). Gaze perception requires focused attention: Evidence from an interference task. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 108–118.

Butts, C. T. (2008). Network: a package for managing relational data in R. *Journal of Statistical Software, 24*(2).

Butts, C. T. (2015). Network: Classes for relational data. https://CRAN.Rproject.org/package=network.

Cabrall, C.D.D., Lu, Z., Kyriakidis, M., Manca, L., Dijksterhuis, C., Happee, R., de Winter, J. (2018). Validity and reliability of naturalistic driving scene categorization Judgments from crowdsourcing. *Accident Analysis & Prevention, 114*, 25–33. https://doi.org/10.1016/j.aap.2017.08.036.

Caird, J. K., Simmons, S. M., Wiley, K., Johnston, K. A., & Horrey, W. J. (2018). Does talking on a cell phone, with a passenger, or dialing affect driving performance? An updated systematic review and meta-analysis of experimental studies. *Human Factors*, *60*(1), 101–133. https://doi.org/10.1177/0018720817748145

Canty, A., Ripley, B., et al. (2012). Boot: Bootstrap R (S-plus) functions. *R package version, 1*(7).

Carsten, O., Hibberd, D., Bärgman, J., Kovaceva, J., Pereira Cocron, M.S., Dotzauer, M., Utesch, F., Zhang, M., Stemmler, E., Guyonvarch, L., Sagberg, F., Forcolin, F. (2017) UDRIVE deliverable 43.1, Driver Distraction and Inattention, of the EU FP7 Project UDRIVE (www.udrive.eu).

Chapman, W. W., Dowling, J. N., & Hripcsak, G. (2008). Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *International Journal of Medical Informatics, 77*(2), 107–113. https://doi.org/10.1016/j.ijmedinf.2007.01.002.

Coxon, A. P. M., & Jones, C. L. (1972). *Multidimensional scaling*. Essex University/European Consortium for Political Research.

Damasio, A. R., Damasio, H., & Van Hoesen, G. W. (1982). Prosopagnosia: anatomic basis and behavioral mechanisms. *Neurology, 32*(4), 331–331.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Vol. 1). Cambridge University Press.

De Swert, K. (2012). Calculating inter-coder reliability in media content analysis using Krippendorff's alpha. *Center for Politics and Communication*, 1–15.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Pereze, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *PNAS, 113*(10), 2636–2641.

Dingus, T. A., Klauer, S.G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Perez, M. A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z. R., Jermeland, J., & Knipling, R.R. (2006) The 100-Car Naturalistic Driving Study, Phase II – Results of the 100-Car Field Experiment, NHTSA, Report No. DOT HS 810 593.

Ekman, F., Johansson, M., Bligård, L.-O., Karlsson, M., & Strömberg, H. (2019). Exploring automated vehicle driving styles as a source of trust information. Transportation Research Part F: Traffic Psychology and Behaviour, 65, 268–279. https://doi.org/10.1016/j.trf.2019.07.026.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251.

Farroni, T., Massaccesi, S., Pividori, D., & Johnson, M. H. (2004). Gaze following in newborns. *Infancy, 5*(1), 39–60.

Fridman, L., Langhans, P., Lee, J., & Reimer, B. (2016) Driver gaze region estimation without the use of eye movement. *IEEE Intelligent Systems, 31*(3), 49–56.

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). IRR: Various coefficients of interrater reliability and agreement. https://CRAN.R-project.org/package=irr.

Hankey, J. M. , Perez, M. A. , and McClafferty, J. A. (2016). Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets. Virginia Tech Transportation Institute, Blacksburg, Va., 2016.

Hayes, A.F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures, 1*(1), 77–89.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour, 1*(10), 743–747.

Hermens, F., Bindemann, M., & Burton, A. M. (2017). Responding to social and symbolic extrafoveal cues: cue shape trumps biological relevance. *Psychological Research, 81*(1), 24–42.

Hermens, F., Kral, D., & Rosenbaum, D. A. (2014). Limits of end-state planning. *Acta psychologica, 148*, 148–162.

Hermens, F., & Walker, R. (2010). Gaze and arrow distractors influence saccade trajectories similarly. *The Quarterly Journal of Experimental Psychology, 63*(11), 2120–2140.

Hermens, F., & Walker, R. (2012). Do you look where I look? Attention shifts and response preparation following dynamic social cues. *Journal of Eye Movement Research, 5*(5):5, 1–11.

Hietanen, J. K., & Leppänen, J. M. (2003). Does facial expression affect attention orienting by gaze direction cues? *Journal of Experimental Psychology: Human Perception and Performance, 29*(6), 1228–1243.

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science, 9*(2), 131–134.

Ioannidou, F., Hermens, F., & Hodgson, T. L. (2017). Mind your step: The effects of mobile phone use on gaze behavior in stair climbing. *Journal of Technology in Behavioral Science, 2*(3-4), 109–120.

Key, C. E. J., Morris, A. P., & Mansfield, N. J. (2016). Situation Awareness: Its proficiency amongst older and younger drivers, and its usefulness for perceiving hazards. *Transportation Research Part F: Traffic Psychology and Behaviour, 40*, 156–168. https://doi.org/10.1016/j.trf.2016.04.011

Kircher, K., & Ahlstrom, C. (2018). Evaluation of methods for the assessment of attention while driving. *Accident Analysis & Prevention, 114*, 40-47. https://doi.org/10.1016/j.aap.2017.03.013.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4*(4), 138–147.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology thousand oaks*. CA: Sage.

Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*.

Lin, R., Liu, N., Ma, L., Zhang, T., & Zhang, W. (2019). Exploring the self-regulation of secondary task engagement in the context of partially automated driving: A pilot study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 147–160. https://doi.org/10.1016/j.trf.2019.05.005

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324.

McNeil, J. E., & Warrington, E. K. (1993). Prosopagnosia: A face-specific disorder. *The Quarterly Journal of Experimental Psychology, 46*(1), 1–10.

Moors, P., Germeys, F., Pomianowska, I., & Verfaillie, K. (2015). Perceiving where another person is looking: the integration of head and body information in estimating another person's gaze. *Frontiers in Psychology, 6*(909). https://doi.org/10.3389/fpsyg.2015.00909

Morando, A., Victor, T., & Dozza, M. (2016). Drivers anticipate lead-vehicle conflicts during automated longitudinal control: Sensory cues capture driver attention and promote appropriate and timely responses. *Accident Analysis & Prevention, 97*, 206–219. https://doi.org/10.1016/j.aap.2016.08.025.

Naqvi, R.A., Arsalan, M., Batchuluun, G., Yoon, H.S., Park, K.R. (2018). deep learning-based gaze detection system for automobile drivers using a NIR camera sensor. *Sensors, 18*(2).

Oram, M. W., & Perrett, D. I. (1994). Responses of anteriorsuperior temporal polysensory (STPA) neurons to biological motion stimuli. *Journal of Neurophysiology, 6*(2), 99–116.

Peng, Y., Boyle, L. N., & Hallmark, S. L. (2013). Driver's lane keeping ability with eyes off road: Insights from a naturalistic study. *Accident Analysis & Prevention, 50*, 628-634. https://doi.org/10.1016/j.aap.2012.06.013.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *The Philosophical Transactions of the Royal Society B: Biological Sciences, 335*, 23–30.

Perrett, D. I., Smith, P. A., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society B: Biological Sciences, 223*, 293–317.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. Bouwhuis (Eds.), *Attention and Performance Vol. X*, Erlbaum (pp. 531–556). Eindhoven, The Netherlands: Institute for Perception Research IPO.

Prati, G., Marín Puchades, V., De Angelis, M., Fraboni, F., & Pietrantoni, L. (2018). Factors contributing to bicycle–motorised vehicle collisions: a systematic literature review. *Transport Reviews, 38(2)*, 184-208. https://doi.org/10.1080/01441647.2017.1314391.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85*(3), 618.

Rayner, K., Pollatsek, A. (1994). *The Psychology of Reading*. Routledge.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PloS ONE, 11*(2), e0150036.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science, 2*, 437–457.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*(2), 252–257.

Rutherford, M. D., & Krysko, K. M. (2008). Eye direction, not movement direction, predicts attention shifts in those with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 38*(10), 1958.

Seppelt, B.D., Seaman, S., Lee, J., Angell, L.S., Mehler, B., & Reimer, B. (2017). Glass half-full: on-road glance metrics differentiate crashes from near-crashes in the 100-car data. *Accident Analysis & Prevention, 107*, 48–62.

Sivak, M., Flannagan, M. J., Miyokawa, T., & Traube, E. C. (2000). Color identification in the visual periphery: consequences for color coding of vehicle signals. *Transportation Human Factors, 2*(2), 135–150.

Swettenham, J., Condie, S., Campbell, R., Milne, E., & Coleman, M. (2003). Does the perception of moving eyes trigger reflexive visual orienting in autism? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1430), 325–334.

Tawari, A., & Trivedi, M. M. (2014). Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos. Paper presented at the 2014 IEEE Intelligent Vehicles Symposium Proceedings.

Tivesten, E., Dozza, M., 2014. Driving context and visual-manual phone tasks influence glance behavior in naturalistic driving. *Transportation Research Part F: Traffic Psychology and Behaviour 26,* 258–272.

van Nes, N., Bärgman, J., Christoph, M., & van Schagen, I. (2019). The potential of naturalistic driving for in-depth understanding of driver behavior: UDRIVE results and beyond. *Safety Science, 119*, 11–20. https://doi.org/10.1016/j.ssci.2018.12.029.

Vora, S., Rangesh, A., & Trivedi, M. M. (2017). On generalizing driver gaze zone estimation using convolutional neural networks. Paper presented at the 2017 IEEE Intelligent Vehicles Symposium (IV).

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences, 15*(4), 160–168.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., Francois, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, H., & Henry, L. (2019). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. https://CRAN.R-project.org/package=tidyr.

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