Visual Attention of Pedestrians in Traffic Scenes: A Crowdsourcing Experiment

Pavlo Bazilinskyy, Dimitra Dodou and Joost C. F. De Winter

Delft University of Technology, Delft, The Netherlands
{p.bazilinskyy, d.dodou, j.c.f.dewinter}@tudelft.nl

Abstract. In a crowdsourced experiment, the effects of distance and type of the approaching vehicle, traffic density, and visual clutter on pedestrians' attention distribution were explored. 966 participants viewed 107 images of diverse traffic scenes for durations between 100 and 4000 ms. Participants' eye-gaze data were collected using the TurkEyes method. The method involved briefly showing codecharts after each image and asking the participants to type the code they saw last. The results indicate that automated vehicles were more often glanced at than manual vehicles. Measuring eye gaze without an eye tracker is promising.

Keywords: Eye gazes · Pedestrians · Automated driving · Crowdsourcing

1 Introduction

Visual attention has been studied extensively from the driver's perspective. To drive safely, a driver needs to allocate attentional resources to relevant elements of the driving task [1], [2]. It is well established that humans examine visual scenes through a combination of task-driven (top-down) and stimulus-driven (bottom-up) attention [3]–[5]. For example, drivers often focus on the vanishing point of the road [3].

The number of pedestrian deaths in traffic fatalities is growing [6]. Most pedestrian casualties occur during road crossing and can be attributed to incorrect crossing decisions [7]. A report by the Netherlands Institute for Road Safety (SWOV) expressed concerns about the lack of knowledge regarding the distractions of non-motorists [8].

Research findings concerning eye movements of drivers cannot necessarily be translated to pedestrians. For example, the notion of the vanishing point is irrelevant for pedestrians, who generally walk at low speed. However, a similar concept of the longest line of sight has been reported to receive pedestrians' attention when navigating in a street environment [9]. Pedestrians also tend to look at the area around the horizon [10].

The crossing decisions of pedestrians depend on the configuration of the traffic scene [11], [12], including the distance to approaching vehicles and traffic density [13]. Pedestrians are more attentive to the road when moving vehicles are present as compared to roads without traffic [14]. In a naturalistic eye-tracking study, Geruschat et al. [15] found that pedestrians' eye and head movements depend on crossing phase (e.g., looking at bollards, curbs, and lines while walking to the curb and looking at cars while standing; turning the head left and right) and crossing strategy (looking at traffic lights when waiting and looking at cars when crossing early). A study with pedestrians walking in a parking garage found that pedestrians' eye movements are directed to specific features, such as the backs of parked vehicles, wheels, the ground, and the area around the driver [16].

Eye trackers tend to be expensive and require participants to come to the laboratory. Fosco et al. [17] used the TurkEyes method for the crowdsourced collection of eye gazes from a browser without specialized hardware. This method assumes that the location of the code reported by the participant corresponds to the last location in the preceding image the participant was looking at.

This study aimed to understand the effect of the distance and type of the approaching vehicle, traffic density, and visual clutter on pedestrians' visual attention. Diverse images of traffic scenes were presented to participants for different exposure times. The experiment was conducted in a crowdsourced setting to collect a large dataset from a diverse population.

2 Method

The research was approved by the Human Research Ethics Committee of the TU Delft. The experiment was conducted via the crowdsourcing platform Appen (https://appen.com). Participants first answered questions on demographics and driving behavior and then proceeded to the experiment, where they looked at images of traffic scenes from the perspective of a pedestrian standing on the pavement and facing the street in front. 107 stimuli were selected based on four parameters:

1. Distance to the most prominent approaching vehicle (car or two-wheeler) [*dist*]: 0 = short (*n* = 60), 1 = medium (*n* = 31), 2 = long (*n* = 16). In 104 images, the approaching vehicle was a car, whereas in 3 images, two-wheelers were labelled as the approaching vehicle. Prominence was assessed subjectively based on the lateral and longitudinal distance to the pedestrian (i.e., driving lane and distance to the crossing line, respectively).
2. Traffic density [*traf*]: 0 = low (one moving vehicle; *n* = 32), 1 = medium (a few moving vehicles; *n* = 38), 2 = high (many moving vehicles; *n* = 37).
3. Visual clutter [*clut*]: 0 = low (only stationary vehicles; *n* = 21), 1 = medium (stationary vehicles and a few other objects; *n* = 60), and 2 = high (stationary vehicles and many other objects; *n* = 26), where objects included people, traffic signs, shop signs, etc. Greenery did not qualify for clutter. Images with a large variety of objects (in terms of type or color) were classified to the high visual clutter category.
4. Type of the approaching vehicle [*veh*]: manual (*veh* = 0; *n* = 81) or automated (*veh* = 1; *n* = 26). Automated vehicles could be recognized by sensory equipment on the roof. Of the 26 images with automated vehicles, 18 depicted various types of Waymo, 6 depicted Uber vehicles, and 2 were a vehicle of Lyft Self-Driving.

The coding was performed manually by the second author. The stimuli were extracted frames from YouTube videos (https://youtube.com), purchased via Shutterstock (https://shutterstock.com), or downloaded from pxfuel (https://pxfuel.com).

For each stimulus, participants were first presented with a fixation cross for 700 ms, followed by the stimulus for 100, 151, 227, 342, 515, 776, 1170, 1762, 2655, or 4000 ms, as defined using a logarithmic scale. There were two groups of participants. In Group 1, participants saw the 107 stimuli only once, with a randomly picked duration for each stimulus. In Group 2, participants were presented with a subset of 10 stimuli for all 10 durations (i.e., 10 stimuli x 10 durations = 100 stimuli). For both groups, the presentation order of the stimuli was randomized. After each stimulus, a codechart with randomly-generated codes linked to coordinates on the stimulus was shown for 700 ms. The participants were then asked to input the code that they saw last. To prevent participants from looking at the at the same area of the screen, four sentinel images requiring the participant to focus on a particular segment on the screen were used (a concept introduced in [17]). Each sentinel image contained an oval with a face.

The experiment was preceded by a training session containing five traffic scene images not included in the experiment and five sentinel images. The participants had to provide at least eight correct codes for the ten images. The participants who failed to complete the training twice were not allowed to participate in the experiment. All images (i.e., fixation cross, stimulus, codechart, sentinel) were 1280x720 px. After completing the experiment, a unique worker code was shown that the participant had to enter to the Appen platform to receive a reimbursement of 0.50 USD.

Heatmaps based on kernel density estimation were created for each stimulus (generated by the seaborn library for Python with default settings [18]). One Area of Interest (AOI) of the most prominent approaching vehicle was manually created for each stimulus on the Image Map Generator (https://image-map.net) by the first author (see Figure 1 for an example). One stimulus contained two two-wheelers approaching simultaneously; these were included in the same single AOI. The surface area of the AOI [*veh_area*] was calculated in pixels for each stimulus. A Pearson correlation matrix of the image characteristics and number of gazes to the AOI was calculated among the 107 images. The script for processing data was created in Python 3.8.5.

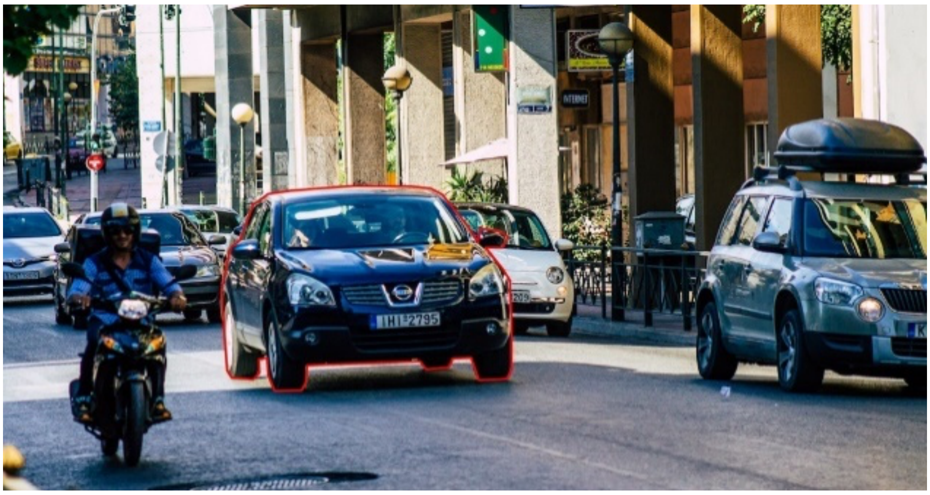


Fig. 1. Example of an AOI around the most prominent approaching vehicle.

3 Results

Between December 3rd 2020 and February 3rd 2021, 3848 attempts to conduct the study were recorded, including people who failed the training session more than twice or did not reach the end of the experiment. 2000 persons completed the experiment. We excluded persons who did not read the instructions ($N = 24$), reported an age under 18 years ($N = 3$), completed the study in less than 5 min ($N = 75$), participated more than once from the same IP address (data from the first run were retained) ($N = 408$), made more than two mistakes with sentinel images ($N = 394$), or provided more than 20% of codechart values with coordinates within a square of 100x100 px around the center of the stimulus ($N = 448$). After filtering, 966 participants remained (mean age 35.5 years, *SD* = 10.4 years; 635 males, 328 females, and 3 participants indicated that they preferred not to respond to the gender question). The mean time for conducting the experiment was 35.3 min (*SD* = 18.3 min). The three most represented countries were Venezuela ($N = 429$), the United States ($N = 86$), and Russia ($N = 71$). The survey was awarded an overall satisfaction rating of 3.7 on a scale from 1 to 5 by 129 participants who completed the satisfaction survey. Groups 1 and 2 contained 839 and 127 participants, respectively.

Figure 2 shows that the distribution of attention varied with exposure time. For this stimulus, the vehicle at a large distance, not in the center of the image, attracted more eye gazes at longer exposure times. Still, participants often focused on the center of the stimulus across all exposure times.

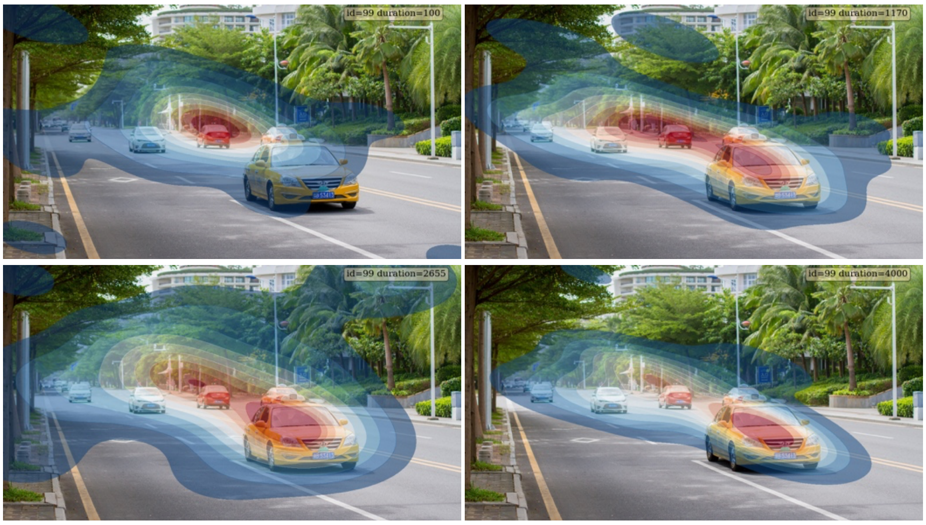


Fig. 2. Exploration of a traffic scene in the stimulus with a vehicle at a large distance, low traffic density, and medium visual clutter for the exposure times of 100, 1170, 2655, and 4000 ms.

Figure 3 shows heatmaps for four selected stimuli aggregated across all exposure times. It is evident that much attention was given to the center of the stimuli.

Figure 4 shows aggregated counts of eye gazes on the object vehicle for all stimuli and all exposure times for all data and split by the four parameters: vehicle distance, traffic density, visual clutter, and vehicle type. It can be seen that the number of gazes to the AOI containing the vehicle plateaued after an exposure time of 342 ms. It can also be noticed that it took time for the participants to detect the vehicle when it was far away (see green bars at the top left figure, representing 5.1% of gazes for 100-ms durations, and 9.1% for 4000-ms durations).

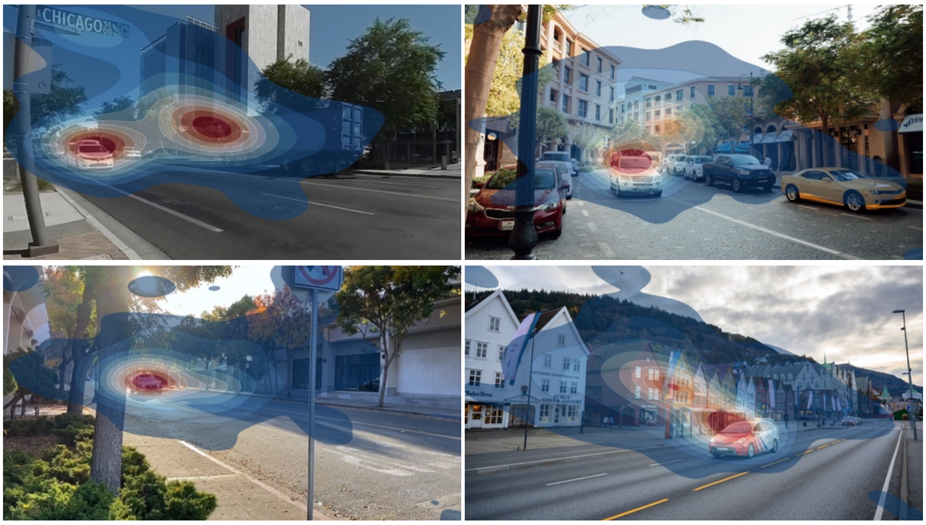


Fig. 3. Heatmaps of four selected stimuli.

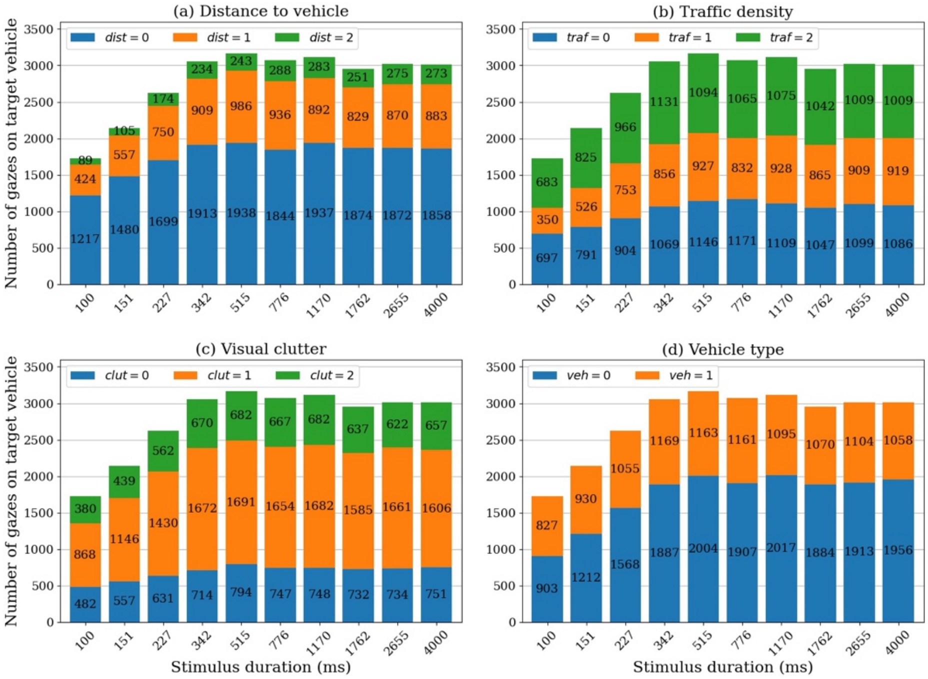
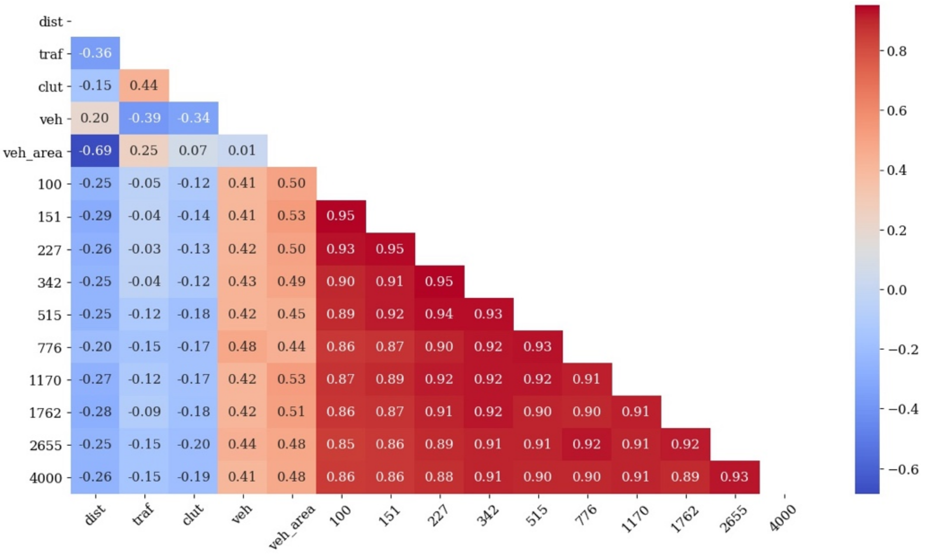


Fig. 4. Aggregated counts of eye gazes on the object vehicle for values of the (a) distance to the vehicle, (b) traffic density, (c) visual clutter, and (d) vehicle type.

Figure 5 depicts, for the 107 stimuli, the correlations of the parameters *dist*, *traf*, *clut*, *veh*, *veh_area*, and the number of participants who gazed at the AOI for each of the ten exposure times. It can be seen that the AOI was less often gazed at in higher traffic density and more cluttered environments. Furthermore, automated vehicles attracted greater attention than manual vehicles, which can be explained by the fact that the former were often presented in low-traffic-density (r = -0.39) and uncluttered (r = -0.34) environments. Finally, larger AOIs, which correspond to more nearby vehicles, attracted more attention.



	dist	traf	clut	veh	veh_area	100	151	227	342	515	776	1170	1762	2655	4000
dist															
traf	-0.36														
clut	-0.15	0.44													
veh	0.20	-0.39	-0.34												
veh_area	-0.69	0.25	0.07	0.01											
100	-0.25	-0.05	-0.12	0.41	0.50										
151	-0.29	-0.04	-0.14	0.41	0.53	0.95									
227	-0.26	-0.03	-0.13	0.42	0.50	0.93	0.95								
342	-0.25	-0.04	-0.12	0.43	0.49	0.90	0.91	0.95							
515	-0.25	-0.12	-0.18	0.42	0.45	0.89	0.92	0.94	0.93						
776	-0.20	-0.15	-0.17	0.48	0.44	0.86	0.87	0.90	0.92	0.93					
1170	-0.27	-0.12	-0.17	0.42	0.53	0.87	0.89	0.92	0.92	0.92	0.91				
1762	-0.28	-0.09	-0.18	0.42	0.51	0.86	0.87	0.91	0.92	0.90	0.90	0.91			
2655	-0.25	-0.15	-0.20	0.44	0.48	0.85	0.86	0.89	0.91	0.91	0.92	0.91	0.92		
4000	-0.26	-0.15	-0.19	0.41	0.48	0.86	0.86	0.88	0.91	0.90	0.90	0.91	0.89	0.93	

Fig. 5. Correlations among stimulus characteristics and number of glances to the vehicle (n = 107).

4 Discussion

In this crowdsourced study, we explored pedestrians' visual attention with a large sample of participants. The focus was on four parameters of a typical traffic scene: (1) distance to the approaching vehicle, (2) traffic density, (3) visual clutter, and (4) type of approaching vehicle.

We found that automated vehicles were more often glanced at than manual vehicles, which can be explained by the fact that images of automated vehicles were often taken from promotional material, with the automated vehicle being depicted in low-traffic-density and uncluttered environments. Moreover, 21 of the 26 depicted automated vehicles were white, which might have attracted attention. We also found that participants required some time to gaze towards the target vehicle. This result is not surprising as participants would need to refocus from the fixation cross to a location on the image and that fixation durations are typically about 250 ms.

A known effect in simulated eye-tracking studies is the center bias [19]. Although care was taken to filter out people who exhibited unrealistic amounts of attention to the central part in the stimuli, our analysis still revealed such bias. This bias is reinforced by the fact that the most prominent vehicle was centered in many of the stimuli.

Crowdsourcing is an efficient data collection method (see [20] for an overview of 15 such studies). In certain cases, results are more robust than data obtained in a comparable lab setting [21]. However, some people sign up for crowdsourced scientific experiments purely because of monetary compensation without paying full attention to given instructions. This project employed a new data filtering method: we rejected work from participants who used the same worker code obtained at the end of experiment more than once. We recommend employing a combination of allocating unique worker codes and rejecting people that reuse them.

We used a computer monitor, offering a limited field of view. Moreover, participants could not move through the environment. Previous research has expressed caution about the use of non-naturalistic data for examining pedestrians' visual attention. Dong et al. [22] found differences in the durations of fixations when navigating or detecting objects in real-world versus simulated environments. The present experiment may be replicated in an on-road setting with the instruction to cross the road.

The dataset in this research may be useful for answering follow-up research questions. For example, it may be used to examine the extent to which pedestrians observe the traffic scene using bottom-up (i.e., looking at salient features) or top-down attention mechanisms (i.e., looking at non-salient but task-relevant features) [3]. The participants in Group 2 were exposed to all durations for each of the presented stimuli. This may make it possible to conduct an autocorrelational analysis. Unlike the large body of literature focusing on the driver, no models of pedestrians' visual attention and exploration of traffic scenes appear to exist. The dataset could be explored for this purpose. In addition, the method may be expanded towards vulnerable road users such as cyclists, perhaps in combination with naturalistic data. Finally, the present dataset could be used to analyze national differences in looking behavior.

Acknowledgement

This research is supported by grant 016.Vidi.178.047 (2018–2024; "How should automated vehicles communicate with other road users?"), which is financed by the Netherlands Organisation for Scientific Research (NWO).

Supplementary Material

Supplementary material with stimuli, their descriptions, and anonymized data are available at https://doi.org/10.4121/13614824. The code used for data processing and analysis is stored at https://github.com/bazilinskyy/gazes-crowdsourced.

References

1. Shinar, D.: Traffic Safety and Human Behavior. Emerald Group Publishing, UK (2017)
2. Lappi, O., Rinkkala, P., Pekkanen, J.: Systematic Observation of an Expert Driver's Gaze

Strategy—An on-Road Case Study. Front. Psychol. 8, 620 (2017)

3. Deng, T., Yang, K., Li, Y., Yan, H.: Where Does the Driver Look? Top-Down-Based Saliency Detection in a Traffic Driving Environment. IEEE Trans. Intell. Transp. Syst. 17, 2051-2062 (2016)
4. Katsuki, F., Constantinidis, C.: Bottom-Up and Top-Down Attention: Different Processes and Overlapping Neural Systems. The Neuroscientist 20, 509-521 (2014)
5. Connor, C.E., Egeth, H.E., Yantis, S.: Visual Attention: Bottom-Up Versus Top-Down. Curr. Biol. 14, R850-R852 (2004)
6. National Highway Traffic Safety Administration: Pedestrian Safety (2018)
7. DaSilva, M.P., Smith, J.D., Najm, W.G.: Analysis of Pedestrian Crashes. Technical report DOT-VNTSC-NHTSA-02-02. National Highway Traffic Safety Administration (2003)
8. Stelling, A., Hagenzieker, M.P. Afleiding in het Verkeer: Een Overzicht van de Literatuur. SWOV Institute for Road Safety Research (2012)
9. Emo, B.: Seeing the Axial Line: Evidence From Wayfinding Experiments. Behav. Sci. 4, 167-180 (2014)
10. Foulsham, T., Walker, E., Kingstone, A.: The Where, What and When of Gaze Allocation in the Lab and the Natural Environment. Vis. Res. 51, 1920-1931 (2011)
11. De Lavalette, B.C., Tijus, C., Poitrenaud, S., Leproux, C., Bergeron, J., Thouez, J.P.: Pedestrian Crossing Decision-Making: A Situational and Behavioral Approach. Saf. Sci. 47, 1248-1253 (2009)
12. Lévêque, L., Ranchet, M., Deniel, J., Bornard, J.C., Bellet, T.: Where Do Pedestrians Look When Crossing? A State of the Art of the Eye-Tracking Studies. IEEE Access 8, 164833-164843 (2020)
13. Rasouli, A., Tsotsos, J.K.: Autonomous vehicles that interact with pedestrians: A survey of theory and practice. IEEE Trans. Intell. Transport. Sys. 21, 900-918 (2019)
14. Tapiro, H., Meir, A., Parmet, Y., Oron-Gilad, T.: Visual Search Strategies of Child-Pedestrians in Road Crossing Tasks. In: De Waard, D., Brookhuis, K., Wiczorek, R., Di Nocera, F., Brouwer, R., Barham, P., Weikert, C., Kluge, A., Gerbino, W., Toffetti, A. (eds.) Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2013 Annual Conference (2014)
15. Geruschat, D.R., Hassan, S.E., Turano, K.A.: Gaze Behavior While Crossing Complex Intersections. Optom. Vis. Sci. 80, 515-528 (2003)
16. De Winter, J., Bazilinskyy, P., Wesdorp, D., De Vlam, V., Hopmans, B., Visscher, J., Dodou, D.: How Do Pedestrians Distribute Their Visual Attention When Walking Through a Parking Garage? An Eye-Tracking Study. Ergon. (2020)
17. Fosco, C., Newman, A., Sukhum, P., Zhang, Y. B., Oliva, A., Bylinskii, Z.: How Many Glances? Modeling Multi-Duration Saliency. In: Workshop on Shared Visual Representations in Human and Machine Intelligence at NeurIPS (2019)
18. Waskom, M.: seaborn.kdeplot-seaborn 0.11.1 documentation (2020) https://seaborn.pydata.org/generated/seaborn.kdeplot.html
19. Tawari, A., Kang, B.: A Computational Framework for Driver's Visual Attention Using a Fully Convolutional Architecture. In: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 887-894. IEEE Press, New York (2017)
20. Bazilinskyy, P., Kyriakidis, M., Dodou, D., De Winter, J.: When Will Most Cars Be Able To Drive Fully Automatically? Projections of 18,970 Survey Respondents. Transp. Res. F: Traffic Psychol. Beh. 64, 184-195 (2019)
21. Bazilinskyy, P., De Winter, J.: Crowdsourced Measurement of Reaction Times To Audiovisual Stimuli With Various Degrees of Asynchrony. Hum. Factors 60, 1192-1206 (2018)
22. Dong, W., Liao, H., Liu, B., Zhan, Z., Liu, H., Meng, L., Liu, Y.: Comparing Pedestrians' Gaze Behavior in Desktop and in Real Environments. Cartogr. Geogr. Inform. Sci. 47, 432-451 (2020)