CooperationCaptcha: On-The-Fly Object Labeling for Highly Automated Vehicles

Marcel Walch, Mark Colley, Michael Weber[*]
Institute of Media Informatics, Ulm University
Ulm, Germany
[firstname.lastname]@uni-ulm.de

ABSTRACT

In the emerging field of automated vehicles (AVs), the many recent advancements coincide with different areas of system limitations. The recognition of objects like traffic signs or traffic lights is still challenging, especially under bad weather conditions or when traffic signs are partially occluded. A common approach to deal with system boundaries of AVs is to shift to manual driving, accepting human factor issues like post-automation effects. We present CooperationCaptcha, a system that asks drivers to label unrecognized objects on the fly, and consequently maintain automated driving mode. We implemented two different interaction variants to work with object recognition algorithms of varying sophistication. Our findings suggest that this concept of driver-vehicle cooperation is feasible, provides good usability, and causes little cognitive load. We present insights and considerations for future research and implementations.

KEYWORDS

Automated driving; human-machine cooperation; study.

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3313022

[*] First and second author contributed equally to this research.

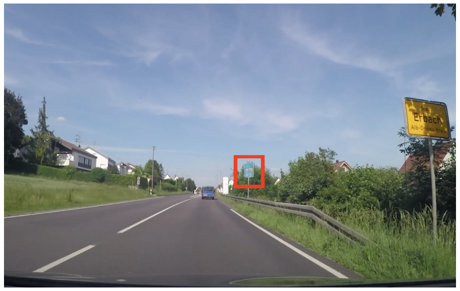
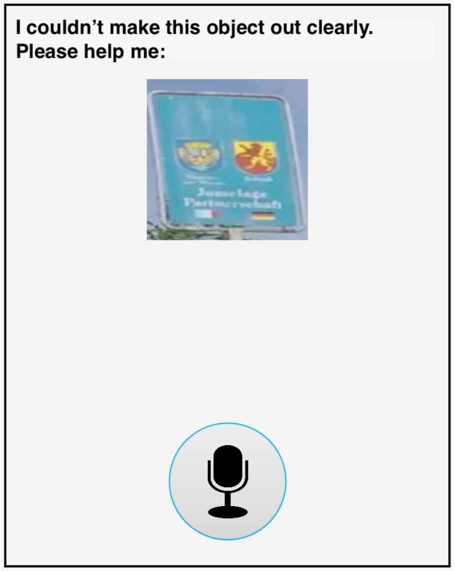


Figure 1: The blue traffic sign on the right (highlighted with a red rectangle) cannot be classified by the system.



I couldn't make this object out clearly.
Please help me:

Figure 2: Free Text System: the system asks for help. After tapping the mic button, the driver can name the sign or in this example "irrelevant" (twin town signage).

INTRODUCTION

Entirely self-driving vehicles that can operate independently under all circumstances are still not within reach [11]. A major challenge for the development of highly and fully automated vehicles is the perception of their surroundings. It is vital that such vehicles recognize other traffic participants as well as road infrastructure such as signalized intersections and road signs, especially when signs are temporary (e.g., due to road works) or electronic/dynamic, and thus not part of the vehicles' high resolution environment map. Real-time object recognition like traffic sign recognition is still challenging at least under bad weather conditions [12]. Factors like occlusion for instance due to snow, tree branches, or graffiti make traffic sign recognition even more difficult. Other recognition challenges can be twisted signs that were hit by another vehicle or informal similar looking signs put up by residents. These issues also apply for recognition of traffic lights and other objects relevant for the driving task. As these issues are easy to solve for humans, we suggest a cooperative approach to overcome this system weakness.

Cooperative driver-vehicle interaction has been proposed as an approach to avoid shifts of control from automated mode to manual driving mode and vice versa [14]. Avoiding handovers (see [7] for a taxonomy of handover situations) means avoiding post-automation usage effects like unstable lateral control [8] or decreased distance to the vehicle in front after platooning [1]. Cooperative driver-vehicle interaction has for instance been implemented to help an automated system choose which action should be conducted next [15]. The strengths of the human driver can be incorporated to fulfill the driving task efficiently without driving manually. Both agents, system and human, become team players and complement each other.

People's superiority over machines regarding the perception and classification of objects has been used in CAPTCHAs to distinguish human users of web sites from bots [3]. Simultaneously, they helped to digitize printed material [13], recognize street names on signs or labeled images [4] like determining where cars are in an image. We suggest to implement this labelling mechanism in AVs. We present the CooperationCaptcha concept: AVs can ask drivers to classify undetected objects to overcome the system's weaknesses on the fly. The benefits of this approach are manifold: labeled training data sets for machine learning, consensus on ambiguously labeled objects, up-to-date map material, transparency regarding system capabilities (calibrated trust [6]) and maintenance of the automated driving mode, thus avoiding handovers of control. Our experiment confirmed that this approach provides good usability and causes low cognitive workload. When participants were provided with possible objects to choose from they were able to label within four seconds. We derived lessons learned and considerations for future work.

ON-THE-FLY OBJECT LABELING

Vehicles that operate in SAE Level 3 [10] (*conditional automation*) require the human driver as fallback. Level 4 automation (*high automation*) does not require the human driver as fallback, however automated driving is only supported in some driving modes [10]. We suggest to implement driver-vehicle

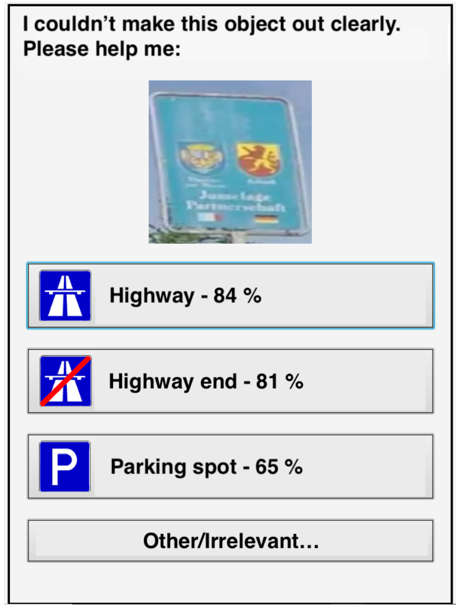

I couldn't make this object out clearly. Please help me:

Figure 3: Choice System: The user can choose among likely traffic signs.

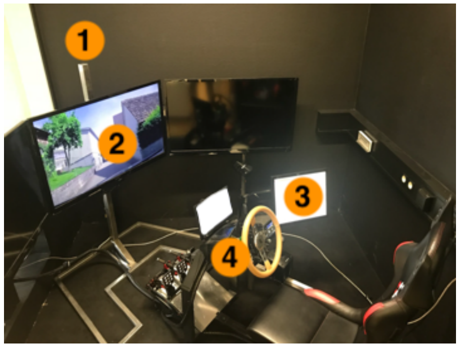


Figure 4: Study setup: camera (1), traffic scene (2), touch screen [GUI] (3) & mic (4).

cooperation to allow conditional and highly automated vehicles to keep the lateral and longitudinal control even in situations where the system is not able to operate entirely on its own due to a lack of information and ultimately to avoid potentially challenging shifts to manual control [1, 8]. Within the scope of this work, we suggest to ask the driver to classify unrecognized objects.

In situations when an AV approaches a scenery where one or several objects are not recognized with enough confidence, the default procedure would be to hand over control from the AV to the human driver. However, such vehicles could ask the human drivers to take responsibility and to classify the objects. We implemented two different interaction techniques for different potential system capabilities. First, we implemented a system that allows the driver to name the unrecognized traffic sign via voice (*free text system*). The second implementation assumes a more sophisticated system that is able to suggest potential traffic signs, for instance based on their color or shape, but is not able to decide which is the correct one. The driver can choose the correct one by means of touch (*choice system*).

Figure 1 shows a scene from the footage used in the experiment. There is a blue sign on the right roadside (twin town signage) that cannot be classified by the vehicle. Consequently, the system announces the need for support with a beep sound and presents a GUI on a screen in the center console. Additionally, the vehicle slows down to indicate uncertainty. Moreover, this increases the time budget for both cooperation partners (system and driver) to classify the object or to plan further actions. The GUI of the free text system is displayed in Figure 2. A screenshot of the choice system providing a selection of potential signs and a *other / irrelevant* button is shown in Figure 3. The free text system also allows the participants to say "other" or "irrelevant" in cases the unrecognized sign is not relevant for the driving task. Moreover, drivers can take over control, for instance in case the choice system does not provide the proper sign. Furthermore, in such cases a combination of both systems could allow the driver to classify the sign anyway.

EXPERIMENT

We conducted a within-subject experiment with 28 participants in a driving simulator to evaluate the basic concept of on-the-fly road sign labeling as well as the two system variants *free text* and *choice*. Participants were engaged in a non-driving related task, considering this very likely in automated driving [9].

Apparatus. Participants sat in the driving simulator displayed in Figure 4. They saw a recording of a real drive through small towns and rural area on the 40" screen in front of the cockpit. When the system asked the participants to label a traffic sign, the playback speed of the recording was reduced during the interaction to increase the time budget and to simulate a decrease of driving speed. The user interface (see Figure 2) was displayed on a 17" touch screen. Moreover, the setup was equipped with a camera to monitor participants and a microphone for voice recording. Speech recognition was

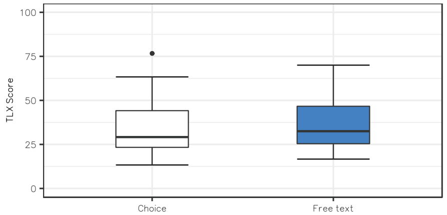


Figure 5: Raw NASA TLX scores

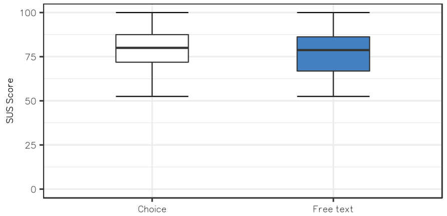


Figure 6: SUS scores

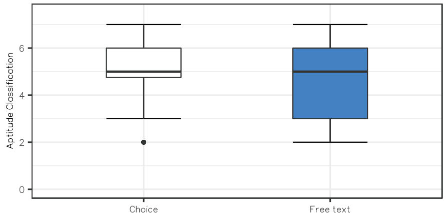


Figure 7: Aptitude scores

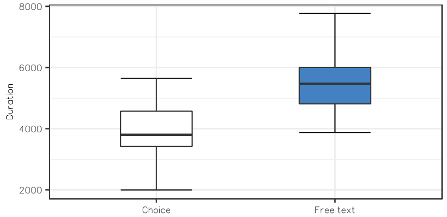


Figure 8: Input duration (in ms)

performed by the experimenter as Wizard of Oz. The Tetris-like game Blockinger* was displayed on an Honor 8 phone running Android 6.0.0 and served as a non-driving related task.

Procedure. After giving consent, participants were introduced to the setup and the seat was adjusted to their needs. It was explained that the system and not the performance of the participant was under evaluation. Moreover, they were told about the purpose of the study and that any questions or feedback were appreciated. For familiarization, the participants were shown a short journey through a small town without any interaction. Participants saw two different journeys (Video A & Video B) using one system (free text or choice) per journey. The assignment of the system to the videos was counterbalanced.

In both conditions 3 objects were relevant for the driving task and 5 were not relevant. Approximately 10 seconds before the vehicle would reach the sign, the system asks the participant. If no selection was performed or the time was up (object was passed) the system would eventually simulate to have recognized the object and continue the automated journey. When the participants were not needed for object labeling, they were engaged in the app Blockinger. They had the option to pause the game. However, only 4 participants did this more than 1 time during their session. After each condition, following questionnaires were filled out: NASA TLX [5] and SUS [2]. At the end of each session they had to rate each system on a custom single-item question regarding system aptitude. All participants received 7 € as compensation.

Participants. We recruited 28 participants mainly from the university population, 11 identified themselves as women and 17 as men. They were on average 25 years ($SD = 9$), held a driving license for cars on average for 7.21 years ($SD = 8.82$) and drove on average 23.56 hours per month ($SD = 49.54$).

RESULTS

Mental Workload. Overall, workload (NASA TLX [5]) was rated moderately low (see Fig. 5). A Wilcoxon signed-rank test showed that both system variants do not differ regarding reported workload (free text: $Mdn = 32.50, IQR = 25.42 - 46.67$; choice: $Mdn = 29.17, IQR = 23.33 - 44.17$), $V = 155, p = 0.421, r = -0.106$.

Usability. Figure 6 shows that both systems were rated highly usable (SUS [2]). As a Wilcoxon signed-rank test revealed, both system variants were rated equally usable (free text: $Mdn = 78.75$, $IQR = 66.88 - 86.25$; choice: $Mdn = 80, IQR = 71.88 - 87.50$), $V = 214, p = 0.168, r = -0.181$.

System Aptitude. Participants had to rate whether the experienced system is suitable for object labeling on a 7-point Likert scale that ranged from 1 (*strongly disagree*) to 7 (*strongly agree*). Figure 7 shows that both systems, free text ($Mdn = 5, IQR = 3 - 6$) and choice ($Mdn = 5, IQR = 4.75 - 6$), were rated as suitable. The ratings were for both systems the same (Wilcoxon signed-rank test), $V = 184, p = 0.569, r = -0.075$.

*https://github.com/vocollapse/Blockinger, Accessed: 4th January 2019

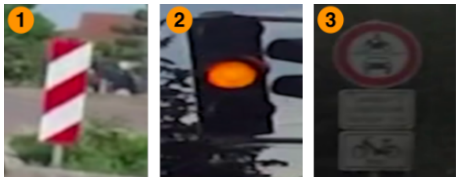


Figure 9: Situations in which participants selected poorly

Video A Relevant	traffic light	traffic light sign	delineator
class	14 of 14	9 of 14	1 of 1
description	11 of 14	13 of 14	0 of 1
relevance	3 of 14	3 of 14	1 of 1

Video B Relevant	traffic light	motor vehicles prohibited	caution children sign
class	13 of 13	4 of 9	1 of 13
description	7 of 13	6 of 9	12 of 13
relevance	1 of 13	0 of 9	1 of 13

	Video A & B Irrelevant
class	39 of 125
description	35 of 125
relevance	90 of 125

Some participants classified irrelevant objects as relevant, which did not render their answer incorrect.

Sidebar 1: Information contained in correct answers. The maximum possible answer count in each cell is 14 (relevant) and 140 (irrelevant).

Input Duration. One participant did not press the microphone button when using the free text system, consequently the according data is excluded from the input duration analysis. Moreover, some interactions were also missing due to going over the time threshold of 7 s to click a button (the object was almost reached after around 7 s) or by saying nothing in the free text condition. Nevertheless, 27 participants contributed a mean interaction duration for the analysis that was calculated of at least 6 interactions. On average, participants needed longer to provide the information when using the free text system ($M = 5428.74$ ms, $SD = 915.4$) than using the choice system ($M = 3966.84$ ms, $SD = 911.94$), Student's paired t-Test: $t(26) = -5.63$, $p < 0.001$, $r = 0.741$. Figure 8 shows the duration of the input in both systems. Participants needed approximately $M = 3.2$ s ($SD = 0.40$) to click the microphone button in the free text condition, but to provide the information they had then to express the information verbally. One word input had a duration of approximately 1 s ("*Sign*") while the longest input measured was about 6.8s ("*Uh a sign where it says what is in the city but not important for the ride*").

Strategy. At the end of each run participants were asked which strategy they had and where their attention was drawn to. 25 of 28 participants stated that they did either not or scarcely look towards the driving scene, both when asked to classify and when no interaction was needed. Their strategy was to look at the picture on the touch screen to perceive the undetected object. This was true for both systems.

Correctness of Classification. With the choice system 87.5 % of the objects were correctly labeled. While the definition of a correct answer in the choice condition is obvious, it is not that clear in the free text condition. We observed answers that only included a statement regarding the relevance of an object (e.g. *"irrelevant"*), only the class of objects (e.g. *"sign"*, but no information regarding the actual name, the meaning of the sign or the state of traffic lights), the name of the sign or a description of the required behavior or action (e.g. *"the car should not go in there"*), or a combination of these (see Sidebar 1). Though, the information given does not consequently suffice for the system to decide what to do, but could be valuable for labeling of the data anyway. There were, however, some situations in which the participants performed poorly with both systems. In these cases contextual knowledge such as location of the object (Figure 9(3)) or temporal change (blinking traffic light, Figure 9(2)) was necessary. Moreover, participants struggled when there were infrequently mentioned signs or many similar-looking signs (Figure 9(1)).

DISCUSSION AND FUTURE WORK

We suggested a cooperative approach to overcome deficiencies in object recognition on the fly to maintain the automated driving mode. This concept has additional benefits such as labelled data to improve object recognition and to update map data. Our preliminary study highlighted the feasibility of the approach: participants reported fairly low workload, gave high usability scores and rated both system variants as suitable for object labeling. Participants were able to select an object from a set of possible objects within four seconds. When participants had to name the unrecognized object via speech they pressed the microphone button after three seconds, however, expressing the information verbally takes extra time which leads to longer input duration. These results show that on-the-fly object labeling can be used to facilitate driver-vehicle cooperation. In turn this can help to avoid handovers [14], and consequently human factor issues of automated driving [1, 8]. These findings constitute a proof of concept and highlight the potential of this approach. Moreover, we derived six insights and considerations for implementations and studies in future research in this area, as described in the sidebar. In future work we will validate our findings in a real-world setting and further improve the interaction design for on-the-fly object labeling.

LESSONS LEARNED AND CONSIDERATIONS

(1) Provide a live preview (video) of the unrecognized object to improve the resolution of the preview and to convey temporal change (e.g. state of traffic lights changes).

(2) Separate *irrelevant* and *other*, since irrelevant objects can be ignored but other objects than the system's proposals can be relevant.

(3) Do not dismiss the dialog as soon as the unrecognized object is passed, since users can still label the object as training data.

(4) Activating the microphone automatically when the dialog appears may reduce input duration.

(5) Speech recognition can result in erroneous inputs that should be correctable by users.

(6) Combining both system variants (showing possible objects and allowing the user to name the object) would combine the advantages of both systems: Faster input when a provided object can be selected and allowing to input the name of the object even when the system does not provide the correct object.

ACKNOWLEDGMENTS

The authors want to thank all study participants. This work was funded by the German Federal Ministry of Education and Research with the funding ID 16SV7624. The authors are responsible for the content of this publication.

REFERENCES

[1] S. Brandenburg and E. M. Skottke. 2014. Switching from manual to automated driving and reverse: Are drivers behaving more risky after highly automated driving?. In 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). 2978–2983. https://doi.org/10.1109/ITSC.2014.6958168

[2] John Brooke et al. 1996. SUS-A quick and dirty usability scale. Usability evaluation in industry 189, 194 (1996), 4–7.

[3] Google Developers. 2018. What is reCAPTCHA? https://developers.google.com/recaptcha/ (accessed 2018–12–27).

[4] Peter Faymonville, Kai Wang, John Miller, and Serge Belongie. 2009. CAPTCHA-based image labeling on the Soylent Grid. In Proceedings of the ACM SIGKDD Workshop on Human Computation. ACM, 46–49.

[5] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology. Vol. 52. Elsevier, 139–183.

[6] Miltos Kyriakidis, Joost CF de Winter, Neville Stanton, Thierry Bellet, Bart van Arem, Karel Brookhuis, Marieke H Martens, Klaus Bengler, Jan Andersson, Natasha Merat, et al. 2017. A human factors perspective on automated driving. Theoretical Issues in Ergonomics Science (2017), 1–27.

[7] Rod McCall, Fintan McGee, Alexander Mirnig, Alexander Meschtscherjakov, Nicolas Louveton, Thomas Engel, and Manfred Tscheligi. 2018. A taxonomy of autonomous vehicle handover situations. Transportation Research Part A: Policy and Practice (2018). https://doi.org/10.1016/j.tra.2018.05.005

[8] Natasha Merat, A. Hamish Jamson, Frank C.H. Lai, Michael Daly, and Oliver M.J. Carsten. 2014. Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. Transportation Research Part F: Traffic Psychology and Behaviour 27 (2014), 274 – 282. https://doi.org/10.1016/j.trf.2014.09.005

[9] Bastian Pfleging, Maurice Rang, and Nora Broy. 2016. Investigating user needs for non-driving-related activities during automated driving. In Proceedings of the 15th international conference on mobile and ubiquitous multimedia. ACM, 91–99.

[10] SAE. 2014. Automated Driving Levels of Driving Automation are Defined in new SAE International Standard J3016. (2014).

[11] Steven E Shladaver. 2016. The truth about "self-driving" cars. Scientific American 314, 6 (2016), 52–57.

[12] Jessica Van Brummelen, Marie O'Brien, Dominique Gruyer, and Homayoun Najjaran. 2018. Autonomous vehicle perception: The technology of today and tomorrow. Transportation research part C: emerging technologies (2018).

[13] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science 321, 5895 (2008), 1465–1468.

[14] Marcel Walch, Kristin Mühl, Johannes Kraus, Tanja Stoll, Martin Baumann, and Michael Weber. 2017. From Car-Driver-Handovers to Cooperative Interfaces: Visions for Driver–Vehicle Interaction in Automated Driving. In Automotive User Interfaces. Springer, 273–294.

[15] Marcel Walch, Tobias Sieber, Philipp Hock, Martin Baumann, and Michael Weber. 2016. Towards Cooperative Driving: Involving the Driver in an Autonomous Vehicle's Decision Making. In Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. ACM, 261–268.