Human Factors in Robots, Drones and Unmanned Systems, Vol. 57, 2022, 146–154
https://doi.org/10.54941/ahfe1002322

Quantifying the Influence of Image Quality on Operator Reaction Times for Teleoperated Road Vehicles

Simon Hoffmann[1], Felix Willert[1], Markus Hofbauer[2], Andreas Schimpe[1], and Frank Diermeyer[1]

[1]Institute of Automotive Technology, Technical University of Munich, 85748 Garching, Germany
[2]Chair of Media Technology, Technical University of Munich, 80333 München, Germany

ABSTRACT

Teleoperated Driving (ToD) is a widely acknowledged concept applied to handle edge-case situations in automated vehicles. In ToD, a human operator judges and resolves these situations based on video streams. Due to varying network coverage, the compression level of these video streams and therefore the resulting image quality (IQ) are adjusted dynamically. In the presented work, the effect of IQ on task performance is investigated. We hypothesize that IQ impacts the operator's reaction time to dynamic obstacles, and therefore influences safety. We conducted a user study to test this hypothesis. Subjective and objective data were collected. The results reveal that IQ has a significant influence on the operator's task performance.

Keywords: Teleoperated driving, Automated driving, Image quality

INTRODUCTION

Motivation

Research on Automated Driving (AD) has made considerable progress in recent years. With SAE Level 4 (SAE International, 2021) or higher, no driver is required inside the vehicle. Hence, no human fallback is inside the vehicle to resolve situations in which the vehicle leaves its Operational Design Domain (ODD). Teleoperated Driving (ToD) is a widely acknowledged concept applied to handle these situations. A human operator connects to the vehicle via a cellular network and resolves the challenging situation remotely. To control a vehicle remotely, the operator requires information about the vehicle's environment to establish situational awareness (Hofbauer, Kuhn, Puttner, et al., 2020; Mutzenich et al., 2021). The operator receives this information primarily via video streams, transmitted from the vehicle to the operator's workstation. The transmission of video streams over a cellular network requires dynamic adaptation of these video streams to compensate for fluctuations in the available transmission rate (Hofbauer, Kuhn, Petrovic, et al., 2020a; Schimpe et al., 2021), leading to varying image quality (IQ). Low IQ can hinder the operator from perceiving the remote environment

correctly, and therefore affect the safety of ToD. This leads to the question: Which IQ is sufficient for safe ToD?

Related Work

Measuring IQ is a wide topic of research. Mohammadi et al. (2014) consider subjective rating to be the most reliable way to quantify IQ. The International Telecommunication Union (2008) provides recommendations to assess subjective IQ. To automate the process of IQ assessment, different objective metrics have been developed. The performance of these metrics is often measured by their correlation with subjective ratings (Pedersen, 2015). However, in some disciplines subjective judgment is less relevant. In radiology, the term task-based assessment of IQ is used to quantify IQ based on the radiologist's task performance (Barrett et al., 2015). This idea also applies to ToD since the main purpose of IQ is to enable safe ToD.

Isozumi et al. (2021) conducted a real driving study to investigate the influence of resolution and frame rate on ToD. The authors determined significant influences on the accuracy of longitudinal and lateral guidance as well as on subjective task performance, difficulty, and fatigue. However, for ToD, IQ is not only influenced by the image resolution, but also by the image compression (Hofbauer, Kuhn, Petrovic, et al., 2020a; Schimpe et al., 2021). Otani et al. (2019) studied the influence of IQ and latency on task-completion time and the number of collisions. Participants had to teleoperate through a labyrinth in a simulated environment. To vary IQ, encoder bitrates between 4.5 Mbps and 1 kbps were used. However, the perceived IQ resulting from a certain bitrate setting depends on the image content (Zhang et al., 2014). Therefore, applying the same bitrate settings to real, non-simulated videos might lead to different results. Georg et al. (2020) investigated the influence of different displays, video canvases and streaming qualities on ToD. They used prerecorded videos and distinguished between the high (15 Mbps) and low quality levels (5 Mbps). This total bitrate is shared across six cameras. Results of a user study indicate a significant influence of IQ on situation awareness (SA), the operator's estimation of the vehicle position within the lane, and the decision making on whether a safe continuation is possible. Neumeier et al. (2020) conducted an online survey to investigate subjective IQ ratings as well as the participants' opinions on whether the provided IQ is sufficient for ToD. Quality levels for the survey were selected using a video quality metric and a clustering algorithm. The authors found a significant influence of driving scenes and quality levels on the subjective rating. Rusák et al. (2014) found significant differences in reaction times between 4K and HD/VGA when playing remote action-reaction games. They assumed that ultra-high resolution provides better visual cues to perceive fine-motoric interactions, concluding that the task has an influence on the measured effect.

From related work, we can conclude that IQ has an influence on ToD. The results of Rusák et al. (2014) indicate that reaction time can be influenced. Since ToD is already prone to latencies (Georg, Feiler, et al., 2020), additional delays pose a potential safety risk. Since Rusák et al. (2014) did not consider image compression and also found a task dependency in their results, we

cannot directly apply these results to ToD. Therefore, we conducted a user study to investigate the influence of IQ on operator reaction times. We hypothesized that operators exposed to lower IQ are slower to react to dynamic obstacles.

USER STUDY

Scenario

To avoid the potential that the task itself influences task performance more than the IQ, the scenario should not vary much. E.g., it might be easier to recognize a vehicle with a poor IQ than a child. Therefore, we require a scenario that is critical under degrading IQ, but also relevant for the operator. The scenario used in this user study is a ball rolling onto the road from an occlusion. This is challenging due to the small obstacle size. In addition, the scenario requires a reaction, even if the ball is not a critical obstacle itself, because a child could follow. Furthermore, the proposed scenario is reproducible and contains a defined event.

Study Design

A within-subject design is used to test the hypothesis. Thus, each subject is exposed to each quality. A total of n streams S are created based on j recordings V and i assigned qualities Q. This results in $n = j \times i$ different videos. To avoid learning effects, each scene is shown at most once per subject, yielding a maximum of j runs per subject. Therefore, each quality could be shown more than once, leading to more reliable results. To reduce the influence of the recording V on the task performance, the assignment of a quality Q to a video V is randomized. Additionally, the order of scenes is randomized, to avoid possible sequence effects. The duration of each scene ranges between 30 s to 60 s. In order to avoid subjects expecting an event becoming more likely with increasing duration of the video, videos without any event were included in the study.

Video Dataset

The videos for the dataset were recorded using a BFSU3-28S5C camera and a F2.8/5mm low distortion lens. The camera is mounted on a ToD test vehicle (Gnatzig et al., 2013). The camera is set to provide images at a resolution of 960x520 pixels and a frame rate of 30 Hz. To vary the IQ, different encoder settings are used. For encoding, GStreamer (GStreamer Team, 2021) and the x264enc plugin are used, since the x264 video encoder is widely used in ToD (Gnatzig et al., 2013; Hofbauer, Kuhn, Petrovic, et al., 2020b). The event-frame is stored together with the encoded video. Because the ball may only be partially visible for a single frame, the ID of the first frame, which shows the ball in full size is used. The following parameters are used for the x264enc plugin:

- *"pass = qual":* For ToD, it is desirable to specify a target bitrate depending on the network coverage. However, with identical bitrate settings, the IQ can vary between or even within a video. To get defined quality levels for

the experiment, the "constant quality" setting is used. If no bitrate limit is specified, GStreamer applies a limit of ~ 2 Mbit/s.

- *"speed-preset = ultrafast":* Speed presets also influence IQ. Since teleoperation is a latency critical application, "ultrafast" is used (Hofbauer, Kuhn, Petrovic, et al., 2020b).

A total of 29 videos were recorded, including six videos that do not contain an event. Three videos are used for training purposes and 20 videos for data collection. Each of the 20 videos V are encoded using five different quality parameters $Q = \{Q20, Q28, Q36, Q44, Q50\}$, which results in a data-set of 109 videos. Since training- and no-event videos are not used for data collection, their IQ is reduced, but not varied between subjects.

Apparatus

For the user study, a Samsung Odyssey C49G94TSSR Monitor with a resolution of 5120 x 1440 pixels is used. The participants were provided with a steering wheel (Fanatec Clubsport Racing Wheel) and pedals (Fanatec Clubsport Pedals V3). The software for the experiment is implemented using the robot operating system (ROS) (Quigley et al., 2009). An input device node reads the operator's inputs and publishes them as part of a ToD software stack (Schimpe et al., 2022). The video viewer node is based on rqt_image_view (Thomas, 2020), which has been extended to output a timestamp when the event-frame is displayed. Furthermore, it is adapted to only show the video stream on a black background. Another node calculates the reaction time t_r of the operator. t_r is stored together with information about the video and the reaction type. We defined the following reaction types: reacted correctly, reacted too early, and did not react.

Data Collection

For each run, the input device data, the timestamp of the event frame, the ID of the video, the quality of the video, and the ID of the event frame are recorded. In addition, t_r of each subject is evaluated. For this purpose, the timestamp of the event frame and the timestamp of the brake input are subtracted. 0.3 % of the maximum pedal travel is considered to be brake input. Since multiple measurements per subject and quality level were conducted, the average reaction time $\overline{t_r}$ for a quality level was calculated per subject.

In addition to the objective data, a subjective assessment of the IQ is performed. A commonly used method for this purpose is to calculate the Mean Opinion Score (MOS) of a subjective Absolute Category Ranking (ACR). First tests have shown that the selected quality parameters for the study might not reach the upper end of the ACR scale (five categories). In order to avoid ratings being too close to each other, a numerical non-categorial ranking (International Telecommunication Union, 2002) is used. The range of ratings R is restricted to $\{R \in \mathbb{N} \mid 1 \leq R \leq 10\}$. Semantic labels for $R = 1$ ("Bad") and $R = 10$ ("Excellent") were provided. In addition, each subject is asked whether the IQ influenced their task performance and whether the provided IQ is sufficient for ToD. For both questions, the same numerical non-categorial ranking is used with labels "Agree" and "Disagree".

Procedure

At the beginning of the study, the subjects were informed about confidentiality of their personal data, followed by a short introduction on ToD. The procedure of the study was explained to familiarize the subjects with the setup. The subjects were asked to keep their foot on the brake, to avoid some subjects varying their foot position and thus additionally influencing the actuation time. The subjects were informed that some videos will not contain an event. Next, three training runs were performed to ensure that the participants understood the task and were familiar with the setup. Each run consisted of the video and reaction measurement followed by video-specific subjective questions. To avoid learning effects, the single stimulus method was applied, meaning that each video was only shown once. After the test runs, the 24 main runs that were used for data collection were performed. Finally, the demographic data was collected. The study took around 35 minutes per subject.

RESULTS

A total of 34 participants (28 male, 6 female) with an average age of $M = 24.65$ years ($SD = 4.94$) participated in the user study. 94 % of the participants were in possession of a driver's license during the time of the user study. The Shapiro-Wilk test indicates significant ($p < 0.05$) non-normality of the dependent variables for certain quality levels. This accounts for objective as well as subjective data. Therefore, to evaluate the influence of IQ, a non-parametric Friedman's ANOVA is conducted throughout this chapter (Field et al., 2012). The Friedman's ANOVA is followed up with a post-hoc analysis using a Wilcoxon signed-rank test with a Bonferroni correction applied.

Subjective Data

Subjective data shows that participants rated the IQ significantly differently for varying quality levels (χ^2 (4) = 120.4, $p < 0.001$). The subjective rating on whether the provided IQ affects task performance also differs significantly (χ^2 (4) = 114.4, $p < 0.001$). Finally, the influence on the subject's opinion whether the provided IQ is sufficient for ToD is significant (χ^2 (4) = 125.0, $p < 0.001$). The results of the post-hoc analysis are depicted in Figure 1. The results for Q20 and Q28 are not significantly different. A further analysis of the video streams showed that for Q20 and Q28 the default bitrate limit of GStreamer was reached and the desired quality levels were overwritten by rate-control. Therefore, the IQ for Q20 and Q28 do not differ considerably, which explains the similar subjective ratings for Q20 and Q28. Throughout each question, the post-hoc analysis shows a significant difference between Q20 and Q36 ($p < 0.001$), as well as between Q28 and Q36 ($p < 0.001$). The effect size for significant comparison is medium ($r > .3$) to large ($r > .5$). Medium effects only appear when comparing Q36 with Q20 and Q28. Effect sizes are estimated according to Fritz et al. (2012).

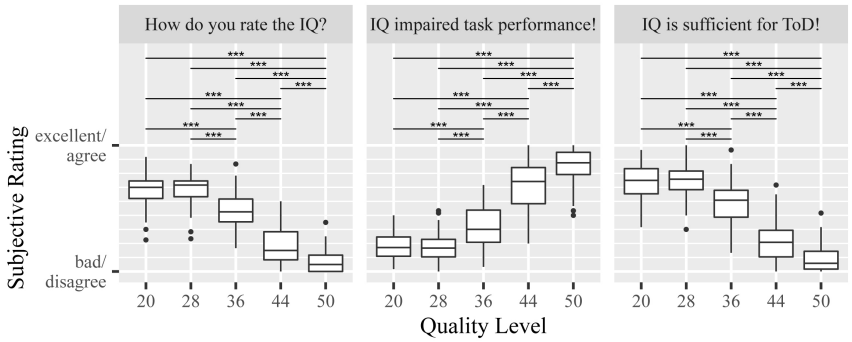

Figure 1: Subjective ratings for different GStreamer quality settings (50: low, 20: high) and the results of a pairwise comparison (***: $p < 0.001$). Q20 and Q28 do not differ considerably, since GStreamer's default bitrate limit was reached for those quality levels.

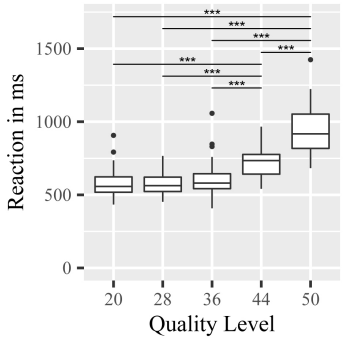

Figure 2: Reaction times for different GStreamer quality settings (50: low, 20: high) and the results of a pairwise comparison (***: $p < 0.001$). Q20 and Q28 do not differ considerably, since GStreamer's default bitrate limit was reached for those quality levels.

Objective Data

The task performance changed significantly over the IQs, (χ^2 (4) = 96.9, $p < 0.001$). The post-hoc analysis indicates that $\overline{t_r}$ at Q50 ($M = 939$ ms) were significantly higher compared to Q44 ($M = 729$ ms, $p < 0.001$, $r = 0.54$), Q36 ($M = 607$ ms, $p < 0.001$, $r = 0.56$), Q28 ($M = 574$ ms, $p < 0.001$, $r = 0.56$) and Q20 ($M = 578$ ms, $p < 0.001$, $r = 0.56$). $\overline{t_r}$ at Q44 were significantly higher compared to Q36 ($p < 0.001, r = 0.47$), Q28 ($p < 0.001$, $r = 0.55$) and Q20 ($p < 0.001$, $r = 0.55$). No significant differences were found between Q20 and Q28, which was expected from the previous section. In contrast to subjective data, there is no significant difference between {Q20, Q28} and Q36. The results of the pairwise comparisons are also highlighted in Figure 2. One participant did not react to the obstacle for a video with quality level Q50. This measurement was excluded from calculating $\overline{t_r}$. Due to multiple measurements per quality level, there was no need to exclude the participant from the user study.

DISCUSSION

The results show significant differences in both subjective and objective data; we can therefore accept the formulated hypothesis. The reaction times of the participants increased significantly for Q44 and Q50. Given the fact that teleoperation is already prone to latencies, these quality levels would definitely decrease driving safety. Q36 is the first quality level for which the subjects perceived a significant difference in IQ compared to higher quality levels {Q20, Q28}. The same applies for subjective task performance and the judgment on whether teleoperation is possible. However, objectively, no significant decrease could be detected for Q36. This could lead to the assumption that the participants are not able to judge the influence on task performance. However, at least for the given scenario, the participants' subjective judgments are more conservative. This, however, is subject to the assumption that an operator would make the correct decision based on their subjective judgment. Using only one scenario might also limit the transferability to road traffic in general. However, the objective was not to find general results for every situation, but to find an upper quality limit using a situation that is most critical in terms of reduced IQ. The participants' feedbacks indicated that even for low qualities the obstacle is easy to detect due to the movement, and that a static obstacle might be more difficult. Therefore, a static obstacle might cause more conservative results.

In addition, further limitations of the study should be emphasized. The measured reaction times cannot be directly transferred to teleoperation in road traffic since the subjects were not driving actively. This has an influence on reaction times as previously investigated by Mackenzie and Harris (2015). However, the scope of this study was not to find absolute reaction times, but whether IQ affects them. Another limitation is that the selected encoder settings (quality levels) for raw images with different resolution also lead to different IQs. Therefore, the results are not valid for raw images with other resolutions. To have a uniform description of the IQs over different settings, IQ metrics can be applied. Their correlation with subjective quality ratings has been well investigated in the literature (Pedersen, 2015), but not for task performance in teleoperation. Furthermore, other parameters, such as brightness and contrast can influence the image quality. However, these influences are independent of the available transmission rate and can be remedied by suitable camera settings or postprocessing.

In a consecutive user-study, a high, user-defined bitrate limit should be set to avoid rate-control and to get data for higher quality levels. Furthermore, two participants did not have a driver's license at the time of the experiment. Since reaction tasks rather than real driving tasks are performed during the experiment, we did not exclude these subjects. Furthermore, participants mentioned dark areas without details in individual videos. However, due to randomization, the overall result is not expected to change.

CONCLUSION

A user study was conducted to investigate the effect of IQ on task performance. We hypothesized that the IQ impacts an operator's reaction time to

dynamic obstacles and therefore influences teleoperation safety. The subjects observed videos of traffic situations and were instructed to brake when necessary. We evaluated the reaction times of the subjects as well as the correctness of their reaction as objective data. Furthermore, the subjective ratings of IQ and the subjective task performances were assessed. The results reveal a significant influence of IQ on the operator's task performance. Subjective data also shows significant differences and indicates that the subjective data is more sensitive to IQ changes than objective data. However, we have discovered it might be more difficult to react to static obstacles or traffic signs. Additionally, the influence of IQ on the SA should be further investigated.

ACKNOWLEDGMENT

Simon Hoffmann initiated the idea for this paper and planned and evaluated the user study. Felix Willert supported the planning and execution of the user study. Markus Hofbauer provided important input on the encoder settings and the planning of the user study. Andreas Schimpe supported the adaption of the encoding pipeline for the data set. Frank Diermeyer made an essential contribution to the concept of the research project. He revised the paper critically for important intellectual content. Frank Diermeyer gives final approval for this version to be published and agrees to all aspects of the work. As a guarantor, he accepts responsibility for the overall integrity of the paper. We acknowledge the financial support for the project by the Federal Ministry of Education and Research of Germany (BMBF) under 16EMO0288.

REFERENCES

Barrett, H. H., Myers, K. J., Hoeschen, C., Kupinski, M. A., & Little, M. P. (2015). Task-based measures of image quality and their relation to radiation dose and patient risk. *Physics in Medicine and Biology*, *60*(2).

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. SAGE Publications Ltd.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1).

Georg, J.-M., Feiler, J., Hoffmann, S., & Diermeyer, F. (2020). Sensor and Actuator Latency during Teleoperation of Automated Vehicles. *2020 IEEE Intelligent Vehicles Symposium*.

Georg, J.-M., Putz, E., & Diermeyer, F. (2020). Longtime Effects of Videoquality, Videocanvases and Displays on Situation Awareness during Teleoperation of Automated Vehicles. *IEEE Intern. Conf. on Systems, Man, and Cybernetics*

Gnatzig, S., Chucholowski, F., Tang, T., & Lienkamp, M. (2013). A system design for teleoperated road vehicles. *2013 10th International Conference on Informatics in Control, Automation and Robotics*, *2*.

GStreamer Team. (2021). *gstreamer - open source multimedia framework*. https://gstreamer.freedesktop.org/

Hofbauer, M., Kuhn, C. B., Petrovic, G., & Steinbach, E. (2020a). Adaptive Multi-View Live Video Streaming for Teledriving Using a Single Hardware Encoder. *2020 IEEE International Symposium on Multimedia*.

Hofbauer, M., Kuhn, C. B., Petrovic, G., & Steinbach, E. (2020b). TELECARLA: An Open Source Extension of the CARLA Simulator for Teleoperated Driving Research Using Off-the-Shelf Components. *2020 IEEE Intelligent Vehicles Symposium*.

Hofbauer, M., Kuhn, C. B., Puttner, L., Petrovic, G., & Steinbach, E. (2020). Measuring Driver Situation Awareness Using Region-of-Interest Prediction and Eye Tracking. *2020 IEEE International Symposium on Multimedia*.

International Telecommunication Union (ITU). (2002). *Methodology for the subjective assessment of the quality of television pictures*.

International Telecommunication Union (ITU). (2008). *Subjective video quality assessment methods for multimedia applications*.

Isozumi, T., Tazaki, Y., Nagano, H., Yokokohji, Y., & Kameoka, S. (2021). Experimental Evaluation of Video Quality Necessary for Remote Driving Follower system. *Society of Automotive Engineers of Japan*, *52*(3).

Mackenzie, A. K., & Harris, J. M. (2015). Eye movements and hazard perception in active and passive driving. *Visual Cognition*, *23*(6).

Mohammadi, P., Ebrahimi-Moghadam, A., & Shirani, S. (2014). Subjective and Objective Quality Assessment of Image: A Survey. *Majlesi Journal of Electrical Engineering*, *9*(1).

Mutzenich, C., Durant, S., Helman, S., & Dalton, P. (2021). Situation Awareness in Remote Operators of Autonomous Vehicles: Developing a Taxonomy of Situation Awareness in Video-Relays of Driving Scenes. *Frontiers in Psychology*, *12*.

Neumeier, S., Stapf, S., & Facchi, C. (2020). The visual quality of teleoperated driving scenarios how good is good enough? *2020 International Symposium on Networks, Computers and Communications*.

Otani, I., Yaguchi, Y., Nakamura, K., & Naruse, K. (2019). Quantitative Evaluation of Streaming Image Quality for Robot Teleoperations. *Artificial Life and Robotics*, *24*(2), 230–238.

Pedersen, M. (2015). Evaluation of 60 full-reference image quality metrics on the CID:IQ. *International Conference on Image Processing*.

Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R., & Ng, A. (2009). ROS: an open-source Robot Operating System. *ICRA Workshop on Open Source Software*.

Rusák, Z., Kooijman, A., Song, Y., Verlinden, J., & Horváth, I. (2014). A study of correlations among image resolution, reaction time, and extent of motion in remote motor interactions. *Advances in Human-Computer Interaction*.

SAE International. (2021). *Surface Vehicle Recommended Practice - J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*.

Schimpe, A., Feiler, J., Hoffmann, S., Majstorovic, D., & Diermeyer, F. (2022). Open Source Software for Teleoperated Driving. *2022 IEEE International Conference on Connected Vehicles and Expo*.

Schimpe, A., Hoffmann, S., & Diermeyer, F. (2021). Adaptive Video Configuration and Bitrate Allocation for Teleoperated Vehicles. *Proc. of Workshop for Road Vehicle Teleoperation at 2021 IEEE Intelligent Vehicles Symposium*.

Thomas, D. (2020). *rqt_image_view* (0..4.16). http://wiki.ros.org/rqt_image_view

Zhang, F., Steinbach, E., & Zhang, P. (2014). MDVQM: A novel multidimensional no-reference video quality metric for video transcoding. *Journal of Visual Communication and Image Representation*, *25*(3).