Comparing the Input Validity of Model-based Visual Attention Predictions based on presenting Exemplary Situations either as Videos or Static Images

Bertram Wortelen (bertram.wortelen1@uni-oldenburg.de)
Cognitive Psychology, C.v.O. University Oldenburg, Ammerländer Heerstr. 69
Oldenburg Germany

Sebastian Feuerstack (feuerstack@offis.de)
Human Centered Design, OFFIS – Institute for Information Technology, Escherweg 2
26121 Oldenburg Germany

Abstract

Functional cognitive models are used to explain observed human behavior. Applying such models to predict behavior requires generalization of the model to be applied in different application domains but also a careful consideration of model input data validity. Visual attention models have already been validated in various domains. But elicitation techniques to collect valid input data that is reproducible by others are still missing. For visual attention prediction model input data is determined mainly based on discussion between experts and individual experience, which is difficult to reproduce. We use a software tool to support input validity. The tool helps users to create attention models. It uses images of the situations that are investigated for stimulating the users to virtually put themselves into these situations. An experiment (n=40) showed that using looping videos instead of static images stimulates imagination in a different way. It has an effect on the models generated by the users and needs careful consideration.

Keywords: Visual Attention; Human Factors; Supervisory Control; Software-supported method; Cognitive Modeling; Safety.

Introduction

Applying cognitive models for predicting human behavior often requires on the one hand expertise in cognitive modeling and on the other hand profound knowledge of the domain for that they are applied. To tackle the former, modeling tools, such as for instance CogTool have been proposed that are based on "zero-parameter models" (i.e. GOMs, KLM and ACT-Simple) (John et al. 2004) that enable human performance prediction based on automatically generated cognitive models. Such approaches can be applied by e.g. a designer to predict human performance for human-machine interface (HMI) design variants.

Most of the cognitive models cannot be generalized and reduced to a predefined fixed set of operators as they depend on experts' inputs to determine valid input parameters. Visual attention prediction models depend for instance on parameters that are knowledge driven and therefore require application domain knowledge for determining the parameter values. The current process for parameter estimation is to ask domain experts and letting them argue and discuss about the parameter values. And in fact studies in various application domains report high correlations to measured data following such approaches (Wickens et al. 2008, Koh et al. 2011). But a discussion-based parameter-determination is hardly reproducible as the quality depends on individual expertise and on the composition of the expert group (to ensure e.g. that silent voices are also heard).

We use a software tool, the Human Efficiency Evaluator (HEE) to capture the relevant knowledge for visual attention prediction in the specific application domain. The tool implements a structured, repeatable process and is used by the experts individually to capture and aggregate their knowledge. To stimulate the knowledge capturing the tool depends on either images or videos that show exemplary situations for that the model parameters are then estimated by the experts.

In this contribution we explore the experts' capabilities to abstract from the very specific concrete shown situations and focus on identifying the differences between a video-based and an image-based stimulus for attention modeling.

Model-based Visual Attention Prediction

Model-based visual attention prediction can complement eye-tracking studies, as it does not depend on HMI functional prototypes and simulations but on human experiences and imagination that is captured by discussion and feed as parameters in prediction models. The SEEV model of attention allocation (Wickens et al. 2001) is such a promising model of visual attention. It describes that "the allocation of attention in dynamic environments is driven by bottom up attention capture of *salient* events, which are inhibited by the *effort* required to move attention, and also driven by the *expectancy* of seeing *valuable* events" (McCarley et al. 2002). The SEEV model is used to predict the percentage of time, that someone spends looking at an area of interest (AOI). It is typically applied by HF experts that have a deep understanding of human attentional processes. The SEEV model relates the probability P_s of attending a specific AOI s to four factors:

$$P_s = \underline{S}aliency - \underline{E}ffort + \underline{E}xpectancy \cdot Task\ \underline{V}alue$$

The first two coefficients, *Saliency* and *Effort* are bottom-up factors that describe the saliency of information displayed by an AOI and the effort it takes to obtain the information, e.g., by moving eyes and head or navigating through a menu. *Expectancy* and *Task Value* are top-down factors. They describe how often new information can be expected

from an AOI and how valuable the information is for accomplishing the tasks of the human operator.

While the bottom up parameters can be estimated e.g. based on physiological data about the effort for eye and head movements (Gore et al. 2009) or by computing saliency maps (Itti & Koch 2001), the determination of the knowledge-based expectancy and value coefficients often depend on data gained by domain experts for a specific application use case.

SEEV model variants, considering some or all of the four factors, have been used to model and predict attention allocations for a wide variety of tasks in various domains: For instance in aeronautics, to predict monitoring while taxiing on ground (Wickens et al. 2008), or the influence of specific cockpit instruments (Goodman et al. 2003) on monitoring behavior. In the automotive domain the model was applied to evaluate drivers' monitoring behavior while approaching intersections (Bos et al. 2015) and also to evaluate the influence of secondary tasks (Wortelen et al. 2013). Recent studies also demonstrate modeling efforts ending with valid predictions for nurses' experience level when assisting in an operation theater in a hospital (Koh et al. 2011). All SEEV model related studies we are aware of, report moderate up to very high correlations ($0.6< R <0.97$) between eye tracking studies and the model predictions.

Improving Input Quality

The broad majority of the studies above applied the "least integer ordinal value" heuristic, which estimates parameter values by letting experts systematically compare AOIs between conditions. A recent approach applies the analytic hierarchy process technique for quantifying the informational importance (Ha & Seong 2014).

The results of those methods, the relevant concrete parameter values are stated in most of the studies above and predictions therefore can be reproduced, but only one study we found (Koh et al. 2011) reported insights about the amount of experts, their background and prior knowledge, and the method applied to agree on the model input parameter values. If the attention model is created for instance by only one HF expert, errors made by this HF expert can have a huge impact on the predictions. If the parameter estimation is a result of a discussion of several experts, quiet voices can be missed easily. Finally, if instead several experts are individually applying a method, the often observed evaluator effect might become evident (Hertzum & Jacobsen 2001).

We use a software tool, the Human Efficiency Evaluator (HEE), for input data gathering. We believe that using a well-structured and tool supported process for input data gathering improves documentation and reproducibility of the input data gathering. Prior studies have shown that the tool can be applied in parallel sessions and with very little training by domain experts for visual attention modeling (Feuerstack & Wortelen 2016). Based on a preset set of operator tasks to consider and images of HMI design variants embedded in their environment, the tool guides the domain experts through four major steps: (1) the identification of areas of interest (AOI) relevant for the operator tasks (see Figure 1 for a screenshot), (2) the determination of expectancy, which is performed by ordering the AOIs according to the expected frequency of information events, (3) ordering the importance of the operator tasks, and finally, (4) the specification of relevance of each AOI for the operator tasks. The least integer ordinal value heuristic is used to calculate numeric parameters from the orders defined in step (3) and (4). In (Feuerstack & Wortelen 2017) we observed a high variance in the data we collected from the domain experts, and interestingly also from the HF experts that we evaluated in a separate session. While variance between experts was also observed in earlier studies e.g. in usability evaluation (Hertzum & Jacobsen 2001) it has not been considered to be relevant for model-based attention prediction to the best of our knowledge. The observed variance seems to be capturing well the diversity that people show in general when asked to give estimates. First studies indicate (Feuerstack & Wortelen 2016, Feuerstack & Wortelen 2017) that the diversity prediction theorem (also called Wisdom of the Crowd (Surowiecki 2004)) can be applied also for attention prediction modeling with the HEE: By averaging individual model predictions, individual prediction errors can be eliminated and high correlations with measured eye-tracking data have been observed (Feuerstack & Wortelen 2017).

To gather expert data the tool requires images representing a situation (e.g. a critical traffic situation) for that the operators' (i.e. drivers') visual attention distribution is then modeled. The approach depends on such images to stimulate the capability of the experts to mentally put themselves into this concrete situation (e.g. one specific left lane change situation) and to anticipate all possible situations that could occur (while performing a lane change). While looking at data from a previous study (Feuerstack & Wortelen 2017), we suspected that the models created by the subjects might be affected by the images that were selected to be representative for a specific situation. Therefore, we investigate how the selection of images representing situations for visual attention modeling impacts the identification of AOIs (i.e. where one looks at) by the subjects and how using videos instead of images might reduce potential biases. For an experiment we formulate the following hypotheses:

H_1: *"Experts mark bigger AOIs for information that is moving relative to the position of the human operator if videos are used to present a driving situation in several variations compared to using static images."*

The location of information that is not fixed relative to the operator is moving in a video, while it has a fixed position in an image. Therefore it is hypothesized that participants only mark boundaries of information at a single position when using images instead of marking larger areas when using videos.

H_2: *"The choice between video and image does not affect the expectancy and value parameters of the SEEV model."*

Although we assume that using videos to represent situations has an effect on the sizes of AOIs compared to using static images, we see no reason, why it should affect the modelling process for expectancy and value parameters of the SEEV model.

H_3: *"More AOIs are marked using looping videos of a situation compared to static images."*

In previous studies (Feuerstack & Wortelen 2016, Feuerstack & Wortelen 2017) we found high individual differences in how many and what kind of AOIs were marked. We assume that the static image is a reason for this variance. Some AOIs might not be marked by every subject, because the dynamics of the situation are not visible in the static image. Thus, they might fail to identify all areas were information shows up. In contrast, the video shows the dynamics of the situation. Thus we assume that more AOIs are marked using videos.

Experiment

We conducted an experiment and asked subjects to model the distribution of attention for different phases of an overtaking maneuver using the Human Efficiency Evaluator (HEE). We tested two conditions in a between subject design with two groups of subjects. For some subjects the driving situations were represented using videos (V condition) and for some using static images (I condition).

Participants

40 licensed car-drivers were recruited by public announcements in the university and were required to be licensed for at least 3 years (mean: 8.05 median: 7.0), have a minimum driving experience of 3000 km per year (mean: 11450 median: 8000) and received an expense allowance of 10 EUR/h. 23 women and 17 men participated in the study, aged between 20 - 40 years (avg: 25.175 median: 24).

Procedure

The experiment was carried out in groups of 4 to 8 subjects for each session and was done in a computer lab in that every subject had a separate PC workplace with two screens. In total we had 20 randomly assigned subjects for each group and participants of both groups were mixed within the sessions.

A video-tutorial, a scripted subject introduction and a written exercise sheet have been used to reduce potential bias by the instructors. Subjects were allowed to ask questions, which were transcribed in the observation records. The subjects had to start with watching the tutorial video first, which introduced them to the tool and its implemented attention modelling process by a supervision example of a football game. The tutorial video was identical for both groups, with the only exception that for one group the football situations were displayed as static images and for the other group looping videos of several variances of the same football situation (a corner kick) were used. After the tutorial were introduced to an overtaking scenario consisting of three phases: (1) merging into left lane, (2) overtaking, and (3) merging into right lane. All subjects were asked to identify all areas of interest for each phase that they assume are relevant for three given tasks as a car driver: (1) Respect speed limit, (2) Overtake slower vehicles, and (3) Control lateral position. Figure 1 depicts the main screen of the HEE that the subjects used to identify the areas of interest. In the video condition the videos started automatically after starting the tool but could be paused by the participants.

After the experiment the two authors and a co-worker independently identified classes of AOIs based on the 1155 AOIs marked by all subjects. In a group discussion we agreed on 37 classes of AOIs. Subjects used for their models different levels of abstraction. For example, some marked the entire dashboard as an AOI, while others differentiated between speedometer, revolution counter and the blinking arrow of the direction indicator. We reflected this by organizing the AOI classes in a hierarchy shown in Figure 2. Afterwards each of the three persons independently classified all 1155 AOIs with a substantial level of agreement (Fleiss' $\kappa = 0.83$).

Results and Discussion

Hypothesis H1

To test hypothesis H1 we differentiate between AOIs that have a fixed position relative to the head of the driver (AOI classes with white boxes in Figure 2), and those that move relative to the head of the driver (AOI classes with gray boxes in Figure 2). As expressed in H1, we only expected an effect for moving information sources.

For each AOI class we took all AOIs belonging to the class, including subclasses and calculated the mean size of the information sources in square pixels.

We did this separately for each condition V and I. The results are plotted in Figure 3. The red line is a straight line

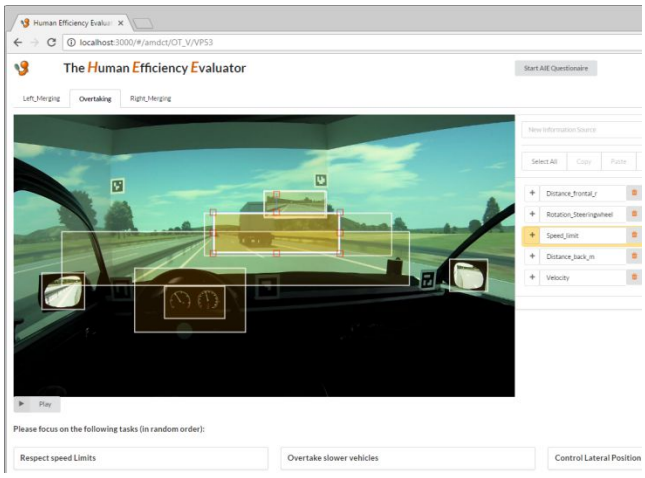

Figure 1: Areas of Interest (AOIs) identification with the HEE.

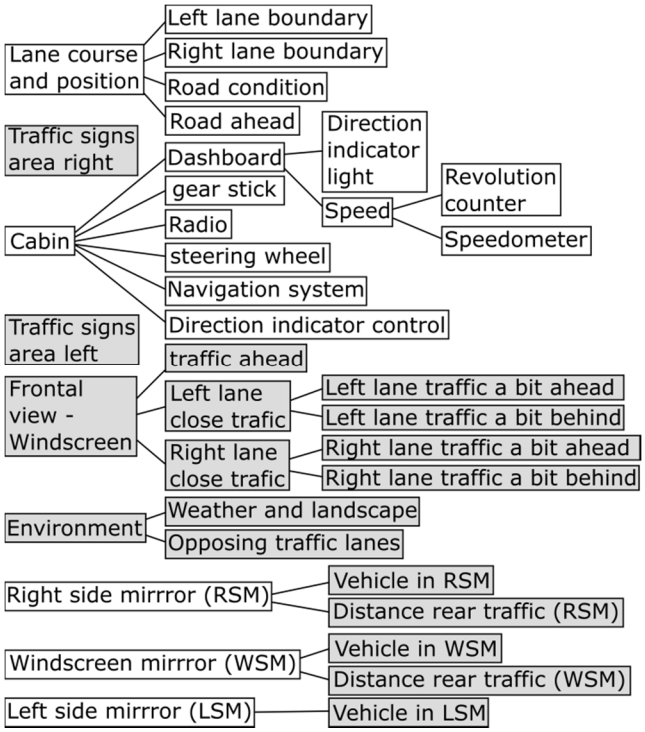

Figure 2: Hierarchy of AOI classes, showing different levels of abstraction.

through the origin with slope 1. For each AOI class the mean AOI size in the V condition is plotted against the mean AOI size in the I condition. According to the H2, AOIs with a fixed position should have similar sizes in both conditions and thus should be located close to the red line (green data points in Figure 3), while moving AOIs should be plotted above the line (blue data points). The sample sizes for all the AOI classes differ, because some AOI classes were marked very often by participants, while others are rarely marked. The size of data points in Figure 3 is proportional to the minimum of the sample size of the V and I conditions (n_{Min}). The Figure seems to support our hypothesis. The two green outliers at (20K, 60K) have a sample size of 1.

For all moving AOI classes with $n_{Min} > 10$, we did Welch two sample t tests with unbalanced sample sizes, to test if the differences are significant. Table 1 shows the p values after Holm-Bonferroni correction for 8 AOI classes with $n_{Min} > 10$. In 4 of the six classes results are significant.

Table 1: Results of the t-tests for differences in AOI sizes between I and V condition for AOI classes with more than 10 AOIs in each condition.

AOI class	p
Frontal view	0.001
Traffic ahead	0.085
Left lane close traffic	0.116
Right lane close traffic	0.003
Right lane traffic a bit ahead	0.011
Traffic signs area right	0.001

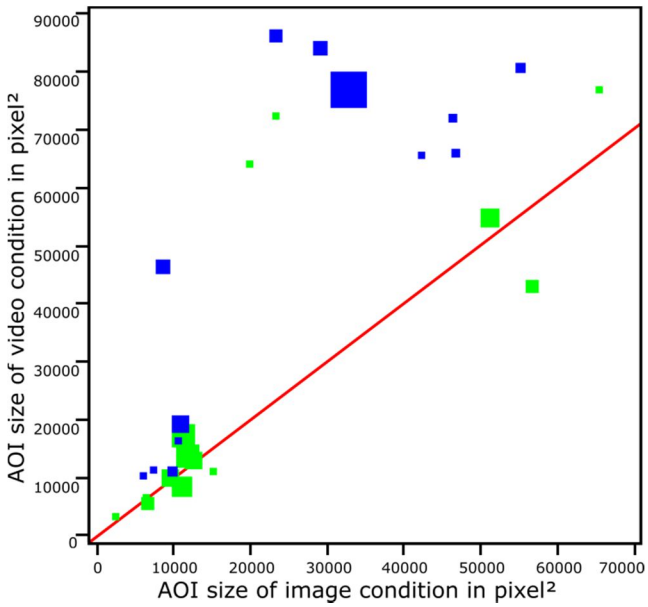

Figure 3. Comparison of average AOI sizes between V and I conditions for each AOI class. AOI classes with moving information are represented with blue data points, fixed AOIs with green data points. Size of data points proportional to sample size. Red line is line through the origin with gradient 1. Blue data points above the line indicate that moving information is marked bigger in the video condition.

Figure 4 illustrates the effect using the AOI class "*Traffic signs area right*" as an example. The top row shows the AOIs marked by subjects of the image group. It mostly shows small areas that just cover the traffic sign in the image, while the areas in the bottom row are from the video condition, were subjects marked the entire region where the traffic sign could be visible. It can also be seen, that participants only mark the information, if it is visible. In the second phase (overtaking) no traffic sign was visible in the image. In this phase only one subject created an AOI in the area were traffic signs are typically perceived.

Hypothesis H2

For testing H2, that there is only an effect on the sizes of AOIs but not on the parameters of the SEEV model, we conducted equivalence tests for these parameters. We used the two one-sided test (TOST) procedure (Schuirmann 1987) to test for equivalence of the parameters between the image and video conditions. For the procedure a margin δ for the difference of the means of the parameters between V and I conditions needs to be defined ($-\delta < \overline{M_V} - \overline{M_I} < \delta$), for which we consider the parameters as equal. [-δ, δ] is the equivalence interval.

Parameters were operationalized using the lowest ordinal heuristic (Wickens et al. 2001). Therefore, the minimum difference between parameters from one modeler is 1. We chose to express the margin δ for the mean of a SEEV parameter for a specific AOI class as a fraction of this minimal individual difference and consider the parameter distribu-

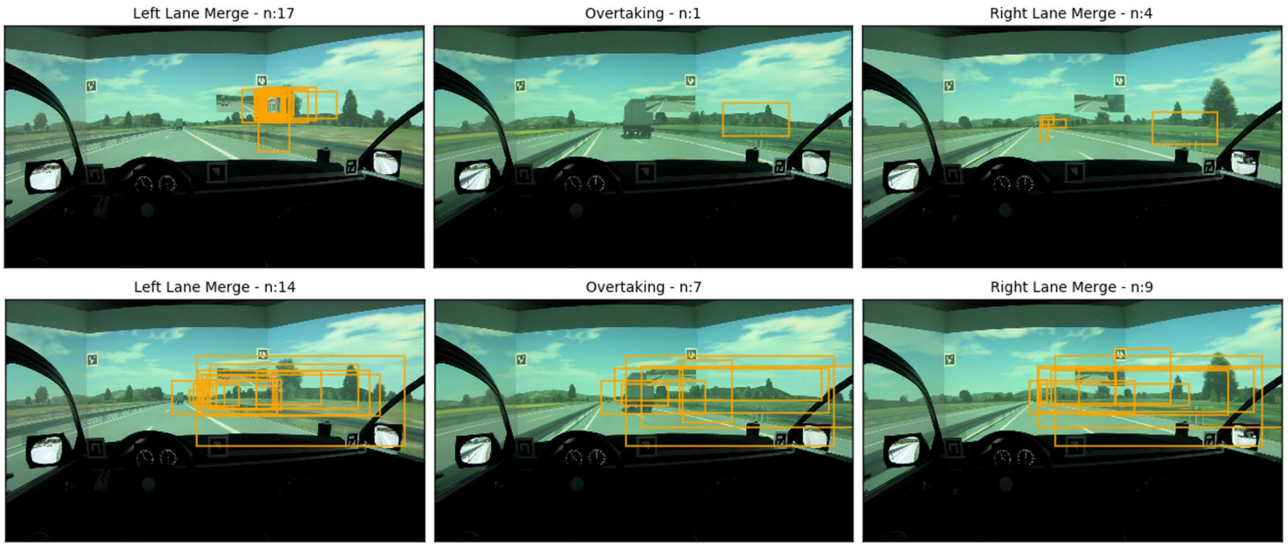

Figure 4: Subjects identification of traffic signs based on static images (top row) and on looping videos (bottom row)

tions equivalent, if the means do not differ more than half of the minimal individual difference (δ = 0.5). The TOST shows equivalence for a chosen α-level, if the (1-2α)-confidence interval is within the equivalence interval. For each AOI class we tested it separately for all three situations, because parameters differ between situations. However, in several cases this resulted in very few data points for an AOI. We excluded AOIs with less than 6 data points.

In Figure 5 all remaining AOIs are listed on the x-axis ordered by the size of the confidence interval. It shows the confidence intervals as red bars and the equivalence interval as blue area. It is easy to see, that we were not able to show parameter equivalence for even a single AOI. For most AOIs the difference of the means is well within the equivalence interval, but the boundaries of the confidence intervals are not. Because we did this test separately for each AOI and each driving situation, the limited number of data points resulted in large confidence intervals and prevents drawing a clear conclusion.

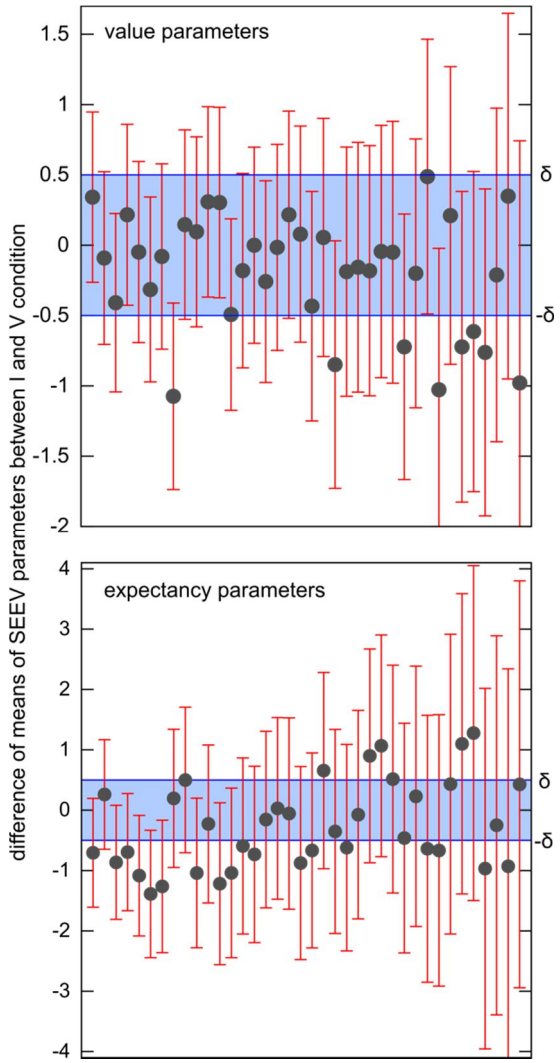

Figure 5. Visualization of the TOST results. Red bars are the 90% confidence intervals. Blue area is the equivalence interval.

Hypothesis H3:

For each participant the identified AOIs for each driving phase were counted resulting in 60=20 × 3 counts. An independent-samples t-test was conducted to compare the counts between video and image condition. There was not a significant difference in the numbers of identified information sources for video (M=6.53, SD=2.13) and image (M=6.23, SD=2.70) conditions (t_{118}=0.68, p=0.50). Subsequent t-tests for each situation alone also found no significant effect.

This result was unexpected. We examined the data in more detail. As we expected, information that is not visible in the image, but is sometimes visible in the video (e.g., indicator lights or road signs) is marked more often in the video compared to the image condition. The opposite case did not occur (information visible in the video, but not in the image). However, we identified another group of AOIs that are visible in the image but only sometimes in the video. This group produces the opposite effect. These AOIs were

marked more often in the image compared to the video condition. A t-test showed, that the difference in the effect was significant (p=0.03) between both groups of AOIs 1: visible in image, but only sometimes in the video, and 2: not visible in image, but at least sometimes visible in video.

Although, this is just a post-hoc hypothesis, it indicates, that the choice between using videos or images and selecting what exactly is displayed has an effect on the models generated by the users. It therefore needs careful consideration.

Conclusion

Exemplary situations to stimulate monitoring behavior modeling need to be carefully chosen. One has to distinguish between areas of information with fixed visual borders (e.g. side mirrors), areas without fixed visual borders but fixed location (e.g. road ahead) from the monitoring person's perspective, and those with moving location (e.g. traffic signs). Specifically the AOI identification of AOIs with moving locations benefits from using videos instead of images.

Acknowledgments

The authors acknowledge the financial support by the European Commission (H2020-MG-2014-2015) in the interest of the project AutoMate – GA 690705 and the funding initiative Niedersächsisches Vorab of the Volkswagen Foundation and the Ministry of Science and Culture of Lower Saxony as a part of the Interdisciplinary Research Centre on Critical Systems Engineering for Socio-Technical Systems.

References

Bos, A. J., Ruscio, D., Cassavaugh, N. D., Lach, J., Gunaratne, P. & Backs, R. W. (2015), Comparison of novice and experienced drivers using the SEEV model to predict attention allocation at intersections during simulated driving, in *Proceedings of the Eighth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*.

Feuerstack, S. & Wortelen, B. (2016), AM-DCT: A visual attention modeling data capturing tool for investigating users' interface monitoring behavior, in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '16, ACM, New York, USA, pp. 252–255. http://doi.acm.org/10.1145/2909132.2909276

Feuerstack, S. & Wortelen, B. (2017), A model-driven tool for getting insights into car drivers' monitoring behavior, in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV'17)* (in press).

Goodman, A., L.Hooey, B., Foyle, D. C. & Wilson, J. R. (2003), Characterizing visual performance during approach and landing with and without a synthetic vision display: A part task study., in D. C. Foyle, A. Goodman & B. L. Hooey, eds, *Proceedings of the 2003 Conference on Human Performance Modeling of Approach and Landing with Augmented Displays*, number NASA/CP-2003-212267, Moffett Field, CA: NASA, pp. 71–89.

Gore, B. F., Hooey, B. L., Wickens, C. D. & Scott-Nash, S. (2009), A computational implementation of a human attention guiding mechanism in MIDAS v5, in V. G. Duffy, ed., *Digital Human Modeling*, Vol. 5620/2009 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 237–246.

Ha, J. S. & Seong, P. H. (2014), 'Experimental investigation between attentional-resource effectiveness and perception and diagnosis in nuclear power plants', *Nuclear Engineering and Design* **278**, 758–772.

Hertzum, M. & Jacobsen, N. E. (2001), 'The evaluator effect: A chilling fact about usability evaluation methods', *Int. J. Hum. Comput. Interaction* **13**(4), 421–443. http://dx.doi.org/10.1207/S15327590IJHC1304_05

Itti, L. & Koch, C. (2001), 'Computational modelling of visual attention', *Nature Reviews Neuroscience* **2**(3), 194–203.

John, B. E., Prevas, K., Salvucci, D. D. & Koedinger, K. (2004), Predictive human performance modeling made easy, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, ACM, New York, NY, USA, pp. 455–462. http://doi.acm.org/10.1145/985692.985750

Koh, R. Y. I., Park, T., Wickens, C. D., Ong, L. T. & Chia, S. N. (2011), 'Differences in attentional strategies by novice and experienced operating theatre scrub nurses', *Journal of Experimental Psychology: Applied* **17**(3), 233–246.

McCarley, J. S., Wickens, C. D., Goh, J. & Horrey, W. J. (2002), A computational model of attention/situation awareness, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 46, SAGE Publications, pp. 1669–1673.

Schuirmann, D. J. (1987), 'A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability', *Journal of Pharmacokinetics and Biopharmaceutics* **15**(6), 657–680. http://dx.doi.org/10.1007/BF01068419

Surowiecki, J. (2004), *The Wisdom of Crowds*, Doubleday.

Wickens, C. D., Helleberg, J., Goh, J., Xu, X. & Horrey, W. J. (2001), Pilot task management: Testing an attentional expected value model of visual scanning, Technical report, NASA Ames Research Center Moffett Field, CA.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M. & Zheng, S. (2008), *Attention-Situation Awareness (A-SA) Model of Pilot Error*, CRC Press/Taylor & Francis Group, chapter 9, pp. 213–239.

Wortelen, B., Baumann, M. & Lüdtke, A. (2013), 'Dynamic simulation and prediction of drivers' attention distribution', *Transportation research part F: traffic psychology and behaviour* **21**, 278–294.