Experimental Validation of an Eye-Tracking-Based Computational Method for Continuous Situation Awareness Assessment in an Aircraft Cockpit

Simon Schwerd[(✉)] and Axel Schulte

Institute of Flight Systems (IFS), University of the German Bundeswehr
Munich (UniBwM), 85579 Neubiberg, Germany
{simon.schwerd,axel.schulte}@unibw.de

Abstract. We propose a continuous situation awareness (SA) assessment method based on eye-tracking and on-line display analysis. In our method, an explicit representation of the pilot's knowledge about the situation is constructed, which contains the pilot's perceived state of the system. To quantify SA for different information relevant to a task, the perceived state is compared to a ground truth to compute a metric of accordance between the pilot's situational picture and the system state. We evaluate this method in a fast jet cockpit simulator experiment with pilots, both professional and in training, from the German Air Force. The participant had to fly a demanding low-level route while reporting threats appearing at random times. We hypothesized, that low accordance between system state and pilot awareness predicts long response time in threat report, low performance in the flight altitude and low subjective SA. Results show, that pilot performance in the low-level flight was correlated to measured average accordance in the aircrafts altitude. Further, subjective SA ratings showed positive association to our measurement results. Response time did not correlate with the measured SA. Apart from that, our measurement approach identified grave decrements for individual participants in the perceptual level of SA, suited for assistance system interventions. We conclude, that the proposed measurement is a promising method to enable assistance systems adapting to the pilot's mental state.

Keywords: Applied cognitive psychology · Human error · Situation awareness · Eye-tracking · Operator state

1 Introduction

Pilot error is one major cause for aviation accidents. Improvements in safety procedures and training can reduce pilot fallibility but cannot eliminate it completely [1]. In this context, adaptive assistance systems could reduce the risk of accidents by removing possible sources of pilot errors.

One cause of pilot error is the loss of situation awareness (SA) [2]. When a pilot is not aware of relevant information due to perceptual errors or divergence in his mental

model, wrong decision-making may result in fatal consequences [3, 4]. With increasing complexity and automation in a cockpit, maintaining a high level of SA is an important aspects of a pilot's work process.

Pilot SA is model of situational cognition, which means that it is affected by the several factors of the situation. It does not result from a single factor, but is influenced by the state of the environment, automation, machine interface and pilot cognition. Therefore, a continuous measure of SA could be valuable input for assistance systems enabling them to adjust to the SA of the pilot. Conventionally, SA is measured for analytical purposes in simulation studies, where the goal is to evaluate operator performance or interface designs. In these cases, freeze-probe methods such as the Situation Awareness Global Assessment Technique (SAGAT) or online query measures such as the Situation Present Assessment Method (SPAM) are used to measure SA [5]. With freeze-probe measures, the simulation is halted at random times, all screens are blanked, and the participant is asked several questions referring to SA relevant information of the operation. On the contrary, online query measures require the participant to answer such queries without interruption of their task. Several reviews and meta-analysis [5–9] discuss the pros, cons and the validity of different measures. Apart from issues described in those studies, these methods cannot be used in an operational setting and therefore, it is not possible to use them as real-time input for adaptive assistant systems.

As an alternative to query methods, eye-tracking might be a way to measure SA in an operational environment [10]. It can be obtained continuously without intrusion and is a strong indicator of pilots' selective attention. Several studies used eye-tracking to correlate eye-tracking parameters (e.g. fixation frequency of relevant display elements) with task performance and query SA measures. In [11], fixation rates, dwell times, and entropy were used as indication for acquisition of information of air traffic controller. It was demonstrated, that controllers with high performance did focus their attention at important times and distributed attention when possible. The authors of [10] implemented an eye-tracking based measurement of SA for a simple monitoring task and compared it to SAGAT. In this case, task performance prediction by eye-tracking was superior to SAGAT. Their results also showed, that the participants attentional distribution depended on the dynamics and state of the task environment, hence SA is situated in the users work process. Based on these results, other studies explored the use of eye-tracking as input for assistant systems. The authors of [12] implemented a cockpit assistant system, which notified a pilot based on a comparison of scan patterns with a visual behavior database. However, evaluation showed mixed results with a lot of room for improvement. In [13], eye-tracking was used to continuously evaluate SA in the supervision of multiple UAVs. It could be demonstrated that this system enhances operator monitoring performance. In [14] and [15], eye-tracking was used to measure pilot workload as input for an adaptive assistance system for pilot helicopters, which showed promising results in operational use.

Most of the related research demonstrates the utility of eye-tracking in lab environments. However, there are three requirements for a measurement method used by an assistance system: First, a statistical approach to predict SA by analyzing eye-tracking data is not enough. While there are certainly meaningful correlations between SA and statistical gaze measures (e.g. fixation frequency on relevant areas), they possess an

averaging nature. Therefore, rare situations of a pilot missing a relevant cue would be ignored when the overall statistical measure is good enough. Thus, statistical indices are missing important outliers, which are the most promising situations for an assistance system intervention. Second, an assistance system requires knowledge about context and the specific information, a pilot is missing. No meaningful information could be drawn from a single number quantifying SA. Third, the method must predict SA continuously and during operation to enable assistance. Most published methods rely on post-processing ([13–15] are one of the few exceptions).

In our contribution, we address these issues by a computational model which does not rely on statistical measures but holds an explicit representation of the information perceived by the pilot.

2   Computational Estimation of SA

The objective of the proposed system is to find potential sources of pilot errors related to the lack of SA. To measure this "lack", three questions are modeled by our approach: What information is relevant? What does the pilot know? What is the state of the system? In the following, we describe how our approach handles each of these questions.
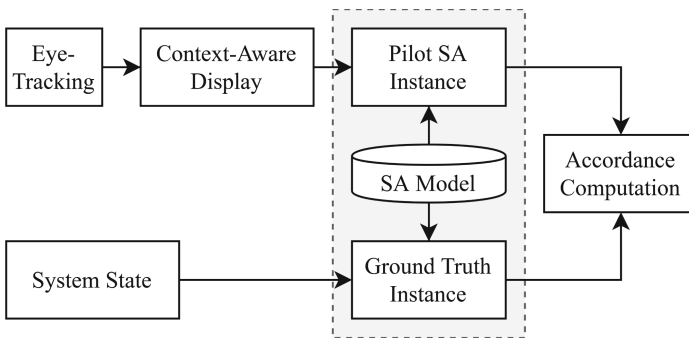

Fig. 1. Overview over the measurement chain

The measurement system is illustrated in Fig. 1. First, the pilot's gaze is measured by an eye-tracking system. The gaze position is used to analyze the fixated display elements in a context-aware interface. This interface application generates information about the display element, the pilot looked at. To infer SA, this information is input to a dynamic knowledge representation. This representation is instantiated from a SA model that associates displayed information to a domain model of SA. The pilot instance of this model is then compared to a similar model fed with the system's ground truth. This comparison is used to compute accordance and identify deviations. The modules are explained more detailed in the following subsections.

2.1 Eye-Tracking and Perception Analysis

The first module provides an estimate of the pilot's visual perception. For this, an eye-tracking system periodically determines the pilot's gaze position in the cockpit. In this stream of gaze positions, only fixated positions are used for further analysis, because they serve as a good indication of information intake [16]. To account for the imprecision of the eye-tracking system, we virtually add gaze samples in a small area around the measured gaze. In the following parts of the measuring chain, these virtual samples are analyzed in similar manner as the measured sample. This ensures that no display element is missed due to small eye-tracking imprecisions. All samples, both measured and virtual, are then passed to the display application.

2.2 Context-Aware Display

The objective of this module is to analyze a gaze sample for the information displayed at that position in the screen. For a given sample, the display application generates a specific observation which contains three types of information: semantic, content and identification. Semantic describes the meaning of the display element at the gaze position, which can be subsequently used to distinguish between all generated observations. The content describes the value of the information at the time of measurement. The identification indicates, which object the information is about.
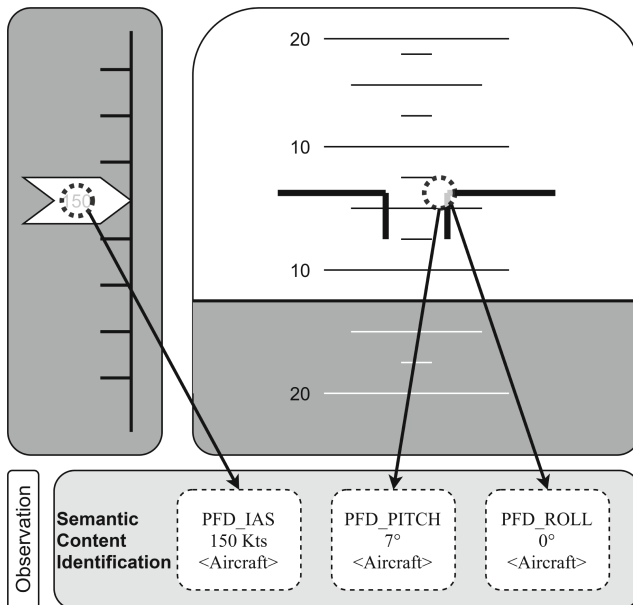

Fig. 2. Simplified primary flight display with two samples

As an example, Fig. 2 shows a simplified primary flight display (PFD) with air speed (left) and attitude (right) indicators: Two hypothetical samples are drawn on air speed indicator and attitude indicator. For the sample on the air speed indicator, the application generates the semantic *PFD_IAS* with a content of $c_{IAS} = 150[kts]$, displayed at measurement. The identification is set to the pilot's aircraft, which is the object, the speed information is about. The other sample on the attitude indicator encodes more than one information alone. It displays both roll and pitch angles of the aircraft by means of an artificial horizon. Therefore, the context-aware display generates two semantics, one for pitch and roll respectively. Both observations contain the value at the time of measurement and a reference to the pilot's aircraft.

2.3   SA Model

To use the continuous stream of observation data as a SA indication, a model is required to associate observations with meaningful information. What an observation *means* is dependent on the application, therefore is domain knowledge. The SA model needs to be flexible enough to capture situational information in different domains. For that purpose, we use a directed logical network. Every node in this network represents a SA-relevant information, subsequently denoted as SA node. The existing observations can be associated to these nodes, which means that a given observation carries the content for the associated SA node. When a gaze measurement generates this observation, the SA node is activated within the network. For example, the *PFD_IAS* from Fig. 2 is an indicator for the aircraft's speed.

As illustrated in Fig. 3, different SA nodes can be connected among each other to represent new aggregations, e.g. the distance between two objects on a map can be estimated by knowing the position of both. This is expressed by edges between SA nodes. These edges implement different logical functions, that define how an initial node activation propagates. However, in this contribution, we only use one type of edge that activates the child node if any parent node is active.
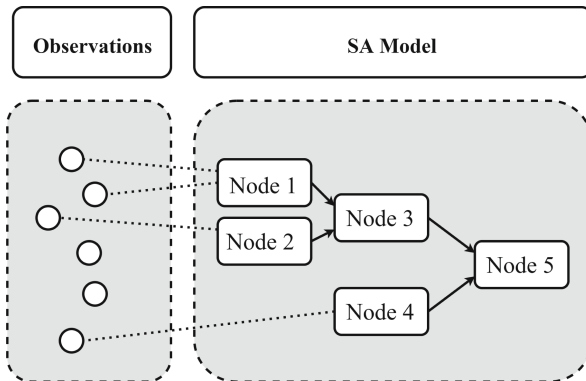

Fig. 3. Situation awareness model

During operation, the SA model is instantiated and fed with the stream of pilot observations. An observation activates every associated node in the model. To account for similar information related to different entities of the same kind, information nodes are uniquely instantiated for every *identification*. This is important, when two display elements represent the same semantic of information but are related to different objects in the environment; e.g. two buildings in a tactical map convey positional information, therefore two nodes are spawned for both buildings respectively. The node activation propagates through the network according to the logical functions of the edges.

A second instance of this model is used for the generation of a ground truth model. For this, we use the system state to generate all possible *observations* in the system containing correct values. These observations are used as input for the SA Model, which represents the ground truth during operation.

2.4 Accordance Computation

The last step is to evaluate the quality of the measured awareness. Thus, we compare the system state to the pilot's awareness. The content of both situational representations – ground truth and pilot's mental state – are used to compute accordance of the pilot's situational picture. In terms of information, a content can be numeric, text, states, times or an identification for another object. To compute the accordance of two values of the same semantic, a parameter c_{norm} *is* required. This parameter quantifies for every observation, what a desirable accordance between two contents is, by normalizing the difference of both. For example, it might be good enough to know the aircrafts altitude within 1000 ft. Note, that this parameter varies for different tasks, e.g. altitude in a low-level flight has a smaller normalizing parameter than a transit flight. For every observation, the accordance is computed as follows:

$$x_{acc} = \begin{cases} 1 - \frac{|c_{pilot} - c_{sys}|}{c_{norm}} & |c_{pilot} - c_{sys}| \leq c_{norm} \\ [3pt] \qquad\qquad 0 & otherwise \end{cases} \tag{1}$$

The terms c_{pilot} and c_{sys} denote the content of the pilot and system observation, respectively. The accordance for text, states and identifications is binary since there is no fuzzy area between two different information of such kind. For numeric values and times, there can be a continuous accordance normalized by c_{norm}. The result of this computational process is a model of the pilot's mental state describing the accordance with the actual ground truth. The accordance is specific to every information node in the system.

2.5 Limitations

For reasons of simplicity, we neglect three phenomena of cognition and perception which might affect SA measurement. First, there is no model for working memory of the pilot. An information which has been fixated upon by the pilot and generated an observation activates all associated nodes in the model. Assuming no change in the content, our model continuously predicts perfect SA without any notion of memory

decay. Second, the perception is solely based on fixations, which are not a perfect measure of visual attention. While there is a high correlation between fixation position and information processing, our approach neglects look-but-failed-to-see phenomena [16]. Third, peripheral vision is neglected for the most part. As described earlier, a small area around a measured fixation is additionally analyzed but apart from that there is no way to observe what the pilot perceives in his peripheral vision.

3   Experimental Evaluation

To evaluate the SA measurement, we conducted a human-machine-experiment. For the experiment, we hypothesized that a low accordance measured by our assessment correlates with low performance as well as low subjective SA. Further, we expected to be able to identify situational errors, which in statistical analysis would be classified as outliers, but represent use-cases for an assistance system.
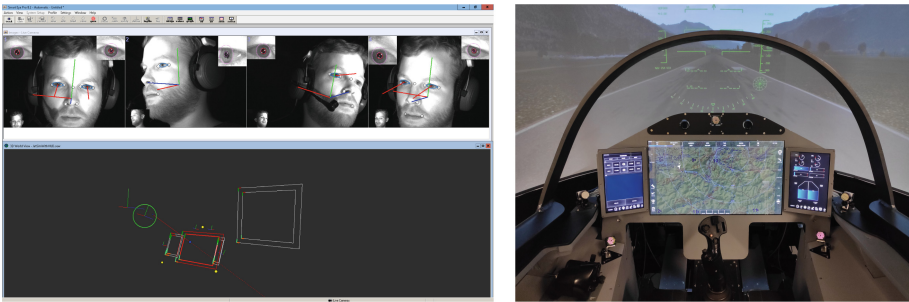

Fig. 4. Integrated gaze tracking system

Flight Simulator. We conducted the experiment in our fast jet research cockpit simulator (Fig. 4). The cockpit does not reflect an existing jet but is a custom research cockpit with three multi-touch-displays. Here, we implemented the measurement chain and fully integrated eye-tracking and the context-aware display to generate observations. For eye-tracking, we used SmartEye©-System connected to our simulator software. The gaze is measured with a sampling frequency of 60 Hz and covers three cockpit displays as well as the Head-Up Display (HUD).

Participants. The experiment was conducted with two groups of participants. The first group consisted of eight professional fighter pilots from the German Air Force with a mean age of 36.25 years (all male, $s_{pilots} = 6.8$). One pilot had to be excluded in the experiment due to data logging issues. The second group were nine pilot candidates from the University of the Bundeswehr Munich with a mean age of 22.6 years (all male, $s_{students} = 1.4$). All participants provided written informed consent.

Procedure. We conducted one experiment with each participant. The two groups received different training. The professional pilots got an extensive training on the simulator with a focus on other experiments. During that time, they did not encounter a

version of the experimental task. They were required to perform the experiment without specific training for the experimental tasks. On the other hand, the pilot candidates were trained for two hours, but did encounter closely related tasks, e.g. low-level flying and threat reporting.

Right before each experiment, the eye-tracking system was calibrated for each participant. Nine-point calibration on every screen resulted in the mean precisions and accuracies given for each participant in Fig. 5. Mean tracking accuracy was $116.2[px]$ and mean standard deviation was $\bar{s}_{acc} = 63.5[px]$. Then, participants of both groups received a short briefing about the upcoming task.
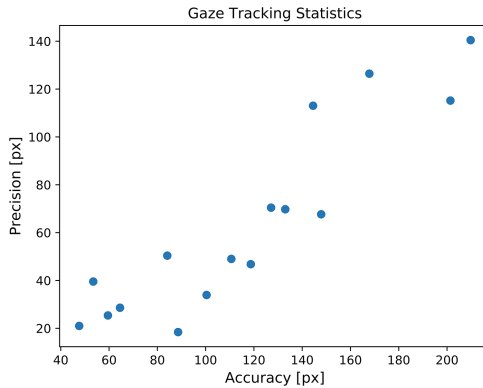

Fig. 5. Gaze tracking statistics

Experiment. In the experiment, the participants had to fly a demanding low-level route through a mountainous environment while keeping their radar altitude below 500 ft. Figure 9 shows an overview of this route. The altitude information was indicated in the HUD radar altitude read-out (marked red in Fig. 6). During the flight, air defenses appeared at random times. No audio cue was given as an indicator of threat. When the participant visually noticed a pop-up air defense either as a red symbol in their central display (Fig. 7) or as a radar warning indication in the HUD (Fig. 6), they had to report the type, their own position and mission time as quickly as possible.
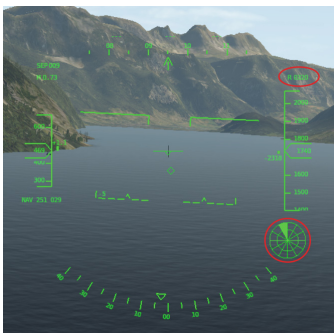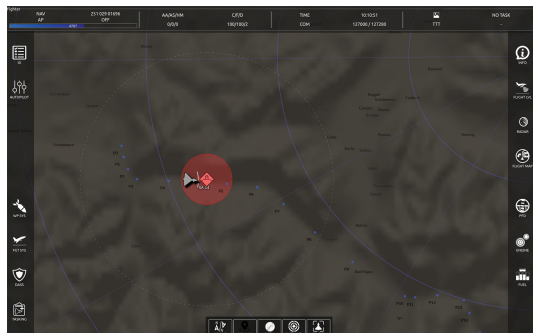

Fig. 6. HUD

Fig. 7. Central cockpit display

As an indirect assessment of SA, we measured performance in two ways. First, we continuously measured if the pilot adheres to the altitude limit. Second, we measured time until the pilot reported his position. The latter can be interpreted as an adapted version of SPAM. The report is not explicitly requested by someone external but is required by an appearing threat. We assumed a quicker threat report to be related to a high pilot accordance in the aircraft position. For subjective SA measurement, the pilots answered a Situational Awareness Rating Technique (SART) questionnaire. For the experiment, we constructed a SA model that describes the relationship among all relevant information for the computational assessment process.
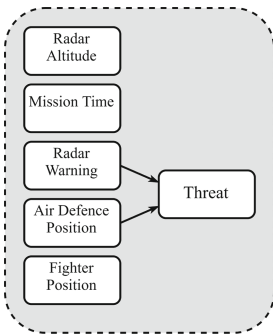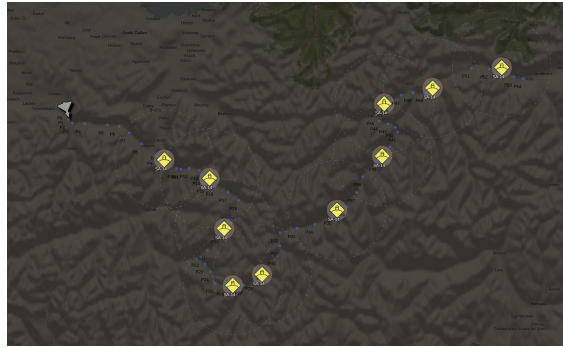

Fig. 8. Experimental model

Fig. 9. Experimental scenario

SA Model. The constructed SA Model, illustrated in Fig. 8, reflects the simplistic nature of the task. The participant had to monitor their radar altitude and look out for a possible threat. The nodes *Radar Altitude* and *Threat* correspond to these tasks in the SA model. In the cockpit, there were two ways to identify a threat: Either in the radar warning display or in a tactical map. In both cases, the node *Threat* is activated for the air defense instance. Note, that a node is created for every new threat since the information relates to different logical instances of the same semantic, which means in this case different enemy air defenses all representing a threat to the pilot. The nodes *Mission Time* and *Fighter Position* are associated with the threat task, where this information must be reported.

Results. We correlated the accordance value of *Fighter Position* of the time right before the threat node was activated with the time duration until the fighter position was reported. We found no correlations between accordance in position and position report time.

Second measure of performance is the aircraft altitude. We assessed accordance of radar altitude to quantify, if low accordance is linked to a pilot violating the altitude limit. For that, we accumulated the time in four categories displayed in a confusion matrix in Table 1. Accordance was classified as high for $x_{acc} > 0.75$ and low otherwise. Note, that we selected a normalizing parameter $c_{norm} = 500$ ft, which means that a deviation of less than 125 ft was classified as high. Altitude was classified as wrong

for $x_{alt} > 500$ ft and correct otherwise. The table contains the mean values and standard deviations of the percentage of time of the combination of two categories over all participants. There is a high percentage of true positives (correct altitude and high accordance), which means that pilot's flying under the limit also had a good knowledge of their altitude. In contrast, there is a great portion of false positives, where the pilots exceeded the altitude limit but, according to the model, were aware of their altitude within the limit. Percentage of mission time in wrong altitude differed strongly between participants with a mean of $\bar{e} = 0.118$ (± 0.118).

Table 1. Confusion matrix for accordance and altitude

	Correct altitude	Wrong altitude
High accordance	.842 (±.0.37)	.618 (±.122)
Low accordance	.128 (±.037)	.382 (±.122)

To study a general link between accordance and error, we computed the correlation between the average accordance of *Radar Altitude* and the time proportion violating the altitude limit. Figure 10 shows the percentage of time violating the altitude limit over the average accordance. Pearson's correlation coefficient is $r = -0.62$ with $p < 0.02$ (two-sided). We excluded one pilot from correlation computation because of his extraordinary bad performance leading to non-normal distributed data. Nevertheless, this outlier fits the trend of a linear regression.
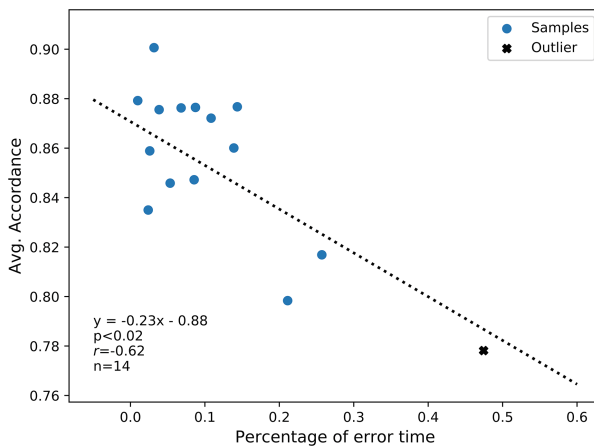

Fig. 10. Scatter plot of error time over average accordance

Further, Fig. 11 shows the average altitude accordance time over the normalized SART score. We computed a positive correlation of $r = 0.56$ with $p < 0.04$ (two-sided). Here, one student participant was excluded due to errors in their SART survey.
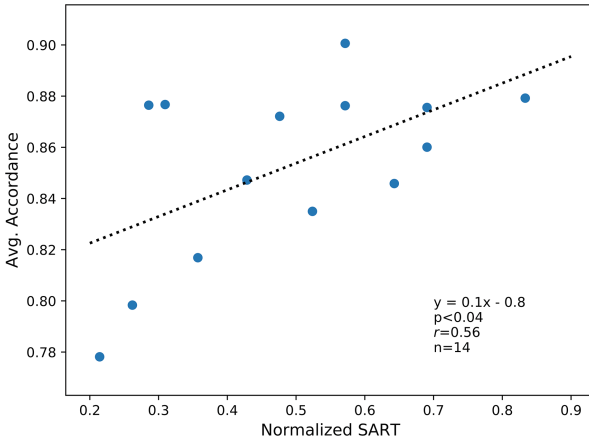
Fig. 11. Scatter plot of the normalized SART score over the average accordance

We could correctly classify the perception of an appearing threat in 73.8% of all samples (n = 16 participants * 10 threats = 160). A threat was classified as correct if the node "*Threat*" was activated and the participant reported within three seconds after activation. In 26.2% of all cases, the threat was not classified correctly due to three different reasons: The pilot reported incorrectly (n = 2), the threat node was inactive when the pilot reported the threat (n = 31) or the threat node was active, but the pilot did not report within three seconds after activation (n = 9).
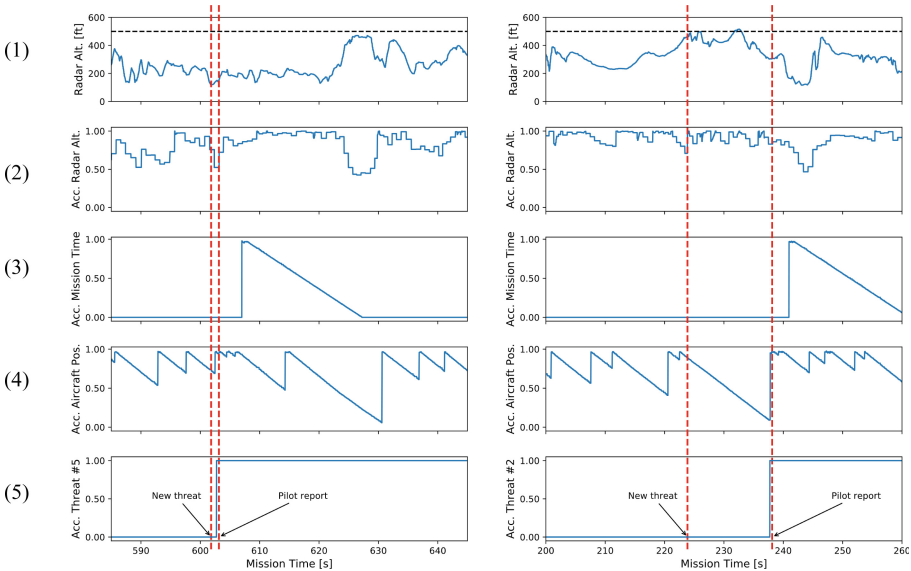

Fig. 12. Accordance for successful report (left) and unsuccessful report (right).

In three cases, a threat was missed for over six seconds by a participant. In these cases, the measurement identified the lack of SA as a deviation in an appearing "*Threat*" node. As an illustrative example, Fig. 12 shows characteristic measurements for both successful and unsuccessful report. The red lines indicate the time of an appearing threat and the time of pilot report, respectively. The data on the left shows, that the pilot quickly notices the threat ($t = 1.7s$), which is indicated in our system by a node activation of *"Threat #2"* (Fig. 12, (5), left) along with a peak in accordance which can be interpreted as the pilot's knowledge about this threat. After the participant starts to report the threat, he updates his knowledge about the mission time as part of the report, which is indicated by an accordance peak in the node *"Mission Time"* (Fig. 12, (3), left). In contrast to that, Fig. 12 on the right shows an extreme outlier of bad performance for the same participant. Here, the pilot misses the threat over a long period ($t = 13.1$ s), which is measured as a late peak in accordance of the threat node (Fig. 12, (5), right). The accordance values of *"Radar Alt"* (Fig. 12, (1), right) indicate, that the pilot was engaged in keeping his altitude below the limit. Nevertheless, the same pattern of updating his knowledge about mission time right after threat report can be observed.

4   Discussion

In this contribution, we sought to evaluate, if our proposed method can assess meaningful aspects of SA. Also, we tried to validate, if this approach can find causes of pilot errors. The results show, that measurement of aspects influencing pilot performance in the experimental tasks was possible. A strong negative correlation of average altitude accordance and error time shows a link between our computational model and performance in the task of low-level flight. Also, the SART results were correlated with the accordance in the flight task. As indicators of pilot errors, statistical outliers such as a pilot missing an enemy air threat were classified correctly. These results indicate, that our system has successfully measured perceptual aspects of SA, which were the dominant factor for performance in the experiment. The confusion matrix shows a high portion of false positives of accordance and error. This is not contradictory to our expectation, since SA can be high when a pilot notices his error. It means that in many cases, a pilot was aware of his altitude while committing an error, which suggests that the participants noticed their error but needed time to adjust. Thus, an adaptive assistance system should respect the pilot working on the solution of a problem.

However, there are also mixed results and limitations of the experimental design. In the threat report task, the response time showed no correlations with measured accordance in the aircrafts position. We assume that response time was not a good SA indicator in our experiment, since some participants were not always eager to answer quickly, and no participant bothered to update his position frequently to answer more quickly. Finally, statistical meaningfulness of the experiment is mitigated by a small number of participants, who were drawn from heterogenous groups (professional vs. unexperienced pilots).

Apart from experimental limitations, the model is based on simplifying assumptions and does not perfectly capture the pilot's perceptual process and mental model,

which was obvious in the look-but-failed-to-see situations of threat report. Also, eye-tracking accuracy and precision differed strongly between participants which had a great influence on prediction quality. Finally, the model does not have a good mechanism to observe pilot intuition about the aircraft's altitude, yet. It is obvious, that spatial information such as altitude can be perceived not only by a numerical sensor indication in the HUD but also by the pilots "feeling" and peripheral vision of the course of the terrain.

5   Conclusion and Further Research

In general, our findings support the idea to assess SA based on eye-tracking and the correlations showed a successful operationalization of SA. Beyond conclusions from statistical inference, our system can provide detailed information about the situational context, which is valuable for an adaptive assistance system. In the proposed method, we addressed requirements for on-line SA measurement listed by [10]: we can measure what the pilot knows, what the system state is and merge this data in a SA model. With this model we can specify relevant information and quantify, how accurate the pilot's SA is.

However, the experiment mainly covered the perceptual level and the SA model was specifically designed for the experimental task. Also, results suffered from eye-tracking measurement errors. Therefore, further research will focus on two aspects: estimation of higher levels of SA by a more complex SA model and increasing measurement accuracy.

References

1. Reason, J.: Understanding adverse events: human factors. Qual. Health Care **4**, 80–89 (1995)
2. Endsley, M., Bolte, B., Jones, D.: Designing for Situation Awareness. CRC Press, Boca Raton (2016)
3. Kelly, D., Efthymiou, M.: An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. J. Safety Res. **69**, 155–165 (2019)
4. Silva, S.S., Hansman, R.J.: Divergence between flight crew mental model and aircraft system state in auto-throttle mode confusion accident and incident cases. J. Cogn. Eng. Decis. Making **9**, 312–328 (2015)
5. Endsley, M.R.: A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. Hum. Factors J. Hum. Factors Ergon. Soc. (2019). https://doi.org/10.1177/0018720819875376
6. Nguyen, T., Lim, C.P., Nguyen, N.D., Gordon-Brown, L., Nahavandi, S.: A review of situation awareness assessment approaches in aviation environments. IEEE Syst. J. **13**, 3590–3603 (2019)
7. Bakdash, J.Z., Marusich, L.R., Cox, K., Geuss, M., Zaroukian, E.: The validity of situation awareness for performance: a meta-analysis. https://psyarxiv.com/kv7n3/. Accessed 24 Jan 2020

8. Kaber, D.B., Prinzel III, L.J., Carolina, N., Prinzel III, L.J., Carolina, N.: Adaptive and adaptable automation design: a critical review of the literature and recommendations for future research. Nasa/Tm-2006-214504 (2006)

9. Salmon, P., Stanton, N., Walker, G., Green, D.: Situation awareness measurement: a review of applicability for C4i environments. Appl. Ergon. **37**, 225–238 (2006)

10. de Winter, J.C.F., Eisma, Y.B., Cabrall, C.D.D., Hancock, P.A., Stanton, N.A.: Situation awareness based on eye movements in relation to the task environment. Cogn. Technol. Work **21**, 99–111 (2018). https://doi.org/10.1007/s10111-018-0527-6

11. van de Merwe, K., van Dijk, H., Zon, R.: Eye movements as an indicator of situation awareness in a flight simulator experiment. Int. J. Aviat. Psychol. **22**, 78–95 (2012)

12. Lounis, C., Peysakhovich, V., Causse, M.: Flight eye tracking assistant (FETA): proof of concept. In: Stanton, N. (ed.) AHFE 2019. AISC, vol. 964, pp. 739–751. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-20503-4_66

13. Fortmann, F., Mengeringhausen, T.: Development and evaluation of an assistant system to aid monitoring behavior during multi-UAV supervisory control: experiences from the D3CoS project. In: ACM International Conference Proceedings Series, pp. 1–8 (2014)

14. Brand, Y., Schulte, A.: Design and evaluation of a workload-adaptive associate system for cockpit crews. In: Harris, D. (ed.) EPCE 2018. LNCS (LNAI), vol. 10906, pp. 3–18. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91122-9_1

15. Honecker, F., Schulte, A.: Automated online determination of pilot activity under uncertainty by using evidential reasoning. In: Harris, D. (ed.) EPCE 2017. LNCS (LNAI), vol. 10276, pp. 231–250. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58475-1_18

16. Newman, B.Y.: Inattentional blindness or looking without seeing. Optom.-J. Am. Optom. Assoc. **82**, 505 (2011)